2014

Computational Study of Protein-Protein Interactions in Misfolded States

Oscar Bastidas
Virginia Commonwealth University

Computational Study of Protein-Protein Interactions in Misfolded States

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science, Chemical and Life Science at Virginia Commonwealth University.

by

Oscar H. Bastidas
Bachelor of Science in Chemical Engineering-Virgina Commonwealth University, 2011

Director: Dr. Michael H. Peters
Professor, Chemical and Life Science Engineering

Virginia Commonwealth University
Richmond, Virginia
August, 2014

Acknowledgement

The author wishes to thank several people. I would like to thank my family for their love, support and patience during the past two years it has taken me to graduate with a Masters degree. I would also like to thank Dr. Peters for his help and for his direction with this project.

TABLE OF CONTENTS

ABSTRACT

Protein-protein interactions (PPI's) play important roles in biological systems. In particular, intra-protein interactions help create and maintain correctly folded protein states and mutations that result in misfolded states may be associated with significant changes in PPI behavior. Six unrelated protein systems with known structure files, each consisting of a wild-type and mutant strain, were studied using the computational algorithm OpenContact©. OpenContact© is a simple tool that can be used to rapidly identify or map interactions "hot-spots" in a protein and was, consequently, used in this study as a starting point to examine the potential or possible role of PPI's on the behavior of mutated, misfolded proteins. Specific results include the observations of single chain protein systems exhibiting mutant strains with significantly stronger inter-atomic interactions as well as a surprising gain of secondary structure in the mutant state. These observations stood in contrast to multi-chain systems (proteins with more than two constituent chains) that appeared to display stronger inter-atomic interactions for the wild-type strains. Results also indicated a potential classification scheme for intra-protein interaction behavior in mutated states based on several criteria. It is important to note, however, that observations on PPI behavior presented need to be verified across a greater number of systems than those studied here before any such trends can be concretely established.

1.0 INTRODUCTION

Proteins are important and extremely vital components of every organism on the planet. Their functions can range from structural support, to membrane transport, metabolism, cell cycle control, and many others. It is therefore not surprising that serious health problems can manifest themselves should there be an aberrant expression of any protein such that its role and function are impaired, or altogether eliminated. Although it is even possible that the novel form of the protein may incur beneficial effects to its host, these events are rare and the more prevalent scenario of aberrant protein expression is the manifestation of disease states. It is therefore beneficial to consider the implications incumbent on mutations that result in misfolded proteins expressed differently from their natural or "wild-type" design. Although mutations of any sort are not to be trivialized, the main class of mutations whose negative effects are to be considered in this work are inherited missense mutations that lead to congenital disease.

By contrast, mutations incurred throughout an organism's lifetime typically affect cell growth and development which can lead to cancer. These two holistic manifestations of disease between inherited and acquired mutations are ultimately different in how they affect the whole organism. Although both can pose life-threatening scenarios, the treatments and considerations for each require different approaches despite the fact that mutations fundamentally lie at the heart of both scenarios. As an example of the unique consequences of genetic damage that can lead to acquired mutations, one can turn to observations in the medical literature that document certain classes of tumors (i.e. ovarian, colorectal, renal, etc…) being associated with genetic damage to specific genes (Gao, Aksoy, Dogrusoz & et al., 2013). For aberrant proteins resulting from inherited mutations, the approach to disease study and pharmaceutical development generally

rests on understanding how the protein's structural differences between the wild-type and mutant strains results in said protein's new bio-chemical and biophysical properties that effectively leads to disease (Jorgensen, 2004). This is the approach that served as the foundation and perspective for the study that was undertaken for this thesis.

2.0 BACKGROUND AND SIGNFICANCE

There exists a plethora of human disorders based on incorrectly folded proteins. Whether one considers disorders as well-known as Cystic Fibrosis and Lou Gehrig's disease or more obscure maladies such as Tay-Sach's disease and Fatal Familial Insomnia, the underlying mode of biological sabotage remains the same: a protein or enzyme is misfolded due to a hereditary missense mutation or an acquired mutation that results in said protein or enzyme either losing its wild-type (and necessary) biological activity or gaining some activity detrimental to biological function (Reynaud, 2010). As alluded to earlier, the importance of proteins in biological systems is immense. In single celled organisms, they can be responsible for anything from membrane transport to digestion (save for autotrophic systems such as euglenas) to even motility (the archetype of which are the cilia of paramecia or the flagella of *Helicobacter pylori*) (Lefebvre, 2001). In viruses, proteins are responsible for reproduction regardless of viral classification (adenoviruses or retroviruses) and most notoriously of all, proteins are essential for host cell entry for all viral classes. In plants, proteins play an important role in embryonic development in spermatophyte species (Soltis, Soltis, & Zanis, 2002). Their importance is even more varied in complex multi-cellular organisms where proteins can play crucial roles in immunization, oxygen transport, structural support and even chemical defense for venomous and poisonous species.

Given the interplay, balance and dependence that many of these systems have with one another for multicellular organisms, one can quickly see that an interruption to any one component can have devastating domino effect-like consequences for the whole organism. Although the results of one system may not directly impact another by providing a bio-chemical product for subsequent use in the next biological process stream, impairing any biological function will place additional strain on any peripheral systems. This strain can be inconvenient at best, or life-threatening at worst. Thus, the importance for quantifying and thoroughly understanding these misfolding events is absolutely necessary to gain appreciation for the states of health and the treatment of disease.

2.1 IN HUMANS: A CLOSER LOOK

In human systems, missense congenital mutations can lead to a variety of disorders that can affect a wide range of life support systems. As an example, Cystic Fibrosis results from a mutation in the gene Cystic Fibrosis Transmembrane Conductance Regulator (CFTR gene) which ultimately leads to an improperly folded version of its respective protein that bears the same name: the CFTR protein (Andersen, 1938). Under wild-type circumstances, this trans-membrane cellular protein is responsible for the regulation of components in sweat and digestive fluids but more importantly, it is responsible for the movement of sodium and chloride ions across the membranes of epithelial cells such as the alveolar epithelia lining of the lungs as well as that of the pancreas and intestines (Andersen, 1938). When a mutation occurs, this function is severely impaired if not eliminated (depending on the actual mutation) and leads to the thick mucus buildup in the lungs characteristic of cystic fibrosis. Cascade events such as pulmonary infections, trouble breathing and poor oxygen delivery can then result complicating matters

further for the patient. As of 2008, the median age of survival in the United States for this disease is 37 years (Andersen, 1938).

Missense mutations that lead to diseases such as cystic fibrosis are not just limited to pulmonary epithelial cells, they can also strike and affect proteins critical to the central nervous system. In Tay-Sach's disease, a mutation on the HEXA gene (which codes for hexosaminidase's, alpha, or "A" subunit) results in a defective version of the enzyme dimer complex, hexosaminidase (the hexosaminidase dimer has an alpha and a beta subunit) (Clarke, Mahuran, Sathe, & et al., 2004). With the alpha or HEXA subunit improperly folded, the entire enzyme is rendered inoperative and incapable of executing its function: degrading ganglioside molecules from propagating ganglion cells of the brain (Clarke et al.). With these molecules rapidly accumulating, they ultimately take on toxic properties that lead to the premature death of the brain's nerve cells (Clarke et al.). The progressive deterioration of these cells results in a decline of mental and physical abilities that can start as early as six months of age in patients suffering from the infantile form the disease (Clarke et al.). Examples of functions that are eventually lost are swallowing and breathing (Clarke et al.). As there is no cure for Tay-Sach's disease, the ultimate prognosis is death which is around four years of age for sufferers of the infantile version but can be variably longer for sufferers of the late onset forms of this disorder (Clarke et al.).

Although the negative effects of misfolding events due to mutations has been strongly emphasized and illustrated up to this point, it is important to keep in mind that not all mutations result in entirely negative consequences for the carrier. A very interesting example that illustrates this is the mutation known as CCR5 delta 32. This mutation results from a deletion of

32 base pairs on a gene known as the CCR5 gene that is normally responsible for expressing the CCR5 protein on the surface of white blood cells such as T cells, macrophages and dendritic cells (Buseyne, Janvier & Teglas, 1998). Under normal circumstances, this cellular surface protein is responsible for serving as a receptor cite for chemokines during immune and immune-related inflammatory responses. When the deletion mutation manifests itself however, the CCR5 protein either fails to be expressed, or is expressed in considerably lower quantities (Buseyne et al., 1998). This has considerable ramifications for viruses that may make use of the CCR5 receptor protein to gain entry into their host cells. HIV is the quintessential virus that wreaks havoc through the CCR5 cellular receptor and those individuals who manifest the mutation either have total immunity to HIV or are able to maintain the viral population at considerably low levels (these specific individuals are known as long-term-nonprogressors or elite controllers) (Buseyne et al., 1998). Due to the inherited genetic nature of this mutation, replicating these altered white blood cells are of active interest in the gene therapy community. The importance of this case serves to establish and emphasize the point that although genetic mutations leading to misfolded or entirely absent proteins can lead to disease, they can confer a genetic advantage if the mutation occurs at the right place and time.

3.0 RESEARCH OBJECTIVES

For the content of this thesis, a total of six misfolded protein cases, or systems, were selected, in part, based on the availability of molecular structure files for both wild and mutant forms of the same protein. Systems containing mutations affecting cysteine residues which would result in the disruption of intra-protein disulfide bridges and, consequently, dramatic/obvious structural changes were purposefully omitted from this study. Each one of the

six cases were identified from the literature and separately analyzed. Each system analyzed was therefore comprised of two protein structure files: a wild type protein structure file and a mutant version of the same protein. As in the foregoing general examples of diseases stemming from protein misfolding, the six systems analyzed for this project also resulted in disease states, but unlike the examples mentioned above, many of the actual systems' respective maladies considered for this project are not as well documented or studied due to the fact that they are not as prevalent or the necessary structure files have only been recently determined.

The overall goal and objective of the research reported here was to identify and quantify the differences in the intra-protein interactions between a wild type protein and a mutant version of that same protein. Consequently, the scope of this project was solely at the intra-protein level. The strategy for determining these differences was via a computational approach based on mapping of the inter-atomic potential field using AMBER 03 force field model. Initially, it was thought that the mutant version of the protein would generally exhibit weaker intra-molecular interactions and a loss of structure compared to the wild type version of the protein. Interestingly, although some systems did exhibit this trend, this was not always the case. Two pieces of software were particularly useful and instrumental in reconciling this information: OpenContact© and DS-Viewer Pro®. OpenContact© provided the computational data and mathematical models surrounding the inter-atomic forces, interactions and distances within each protein that was analyzed. DS Viewer Pro in turn provided qualitative structural visualization images for each protein. In all, data for each system thus consisted of numerical data for the binding parameters analyzed as well as visual molecular images detailing the differences in tertiary and quarternary structure between each case's wild-type and mutant strains.

3.1 MOLECULAR MECHANICS: FOUNDING PRINCIPLES

The computational criterion for quantitatively determining the interatomic attractions and forces is founded on the discipline of molecular mechanics. Molecular mechanics is an application of classical mechanics (or Newtonian mechanics, the most common branch of physics) to model the movement of the individual atomic constituents in molecular systems (Kuhn, Kollman, & et al., 2000). This discipline is used to study a wide range of chemical species behavior ranging from small organic molecules containing only a few atoms, to large biological macromolecules such as proteins and DNA which may contain anywhere from several thousand to several million atoms. As alluded to earlier, the calculations that provided the quantitative data for this project were ultimately executed by OpenContact©, a computer algorithm, but the fundaments of the underlying code are all based upon what follows in this section and needless to say, the calculations are performed by the algorithm for as many atoms that are present in the system being analyzed.

In its most basic manifestation, molecular mechanics models systems via an "all-atom" approach: each individual atom in the molecule is treated as a single particle (Kuhn et al., 2000). Each particle in turn, is then assigned a radius (most typically the van der Waals radius), along with a variety of different parameters that account for polarizability, constant net charge and bond angles (more on these parameters later) (Kuhn et al., 2000). The bonded interactions are modeled as stiff mechanical springs with an equilibrium distance equal to the experimental or calculated bond length (Kuhn et al., 2000). As will be explored in greater detail shortly, the particular force field parameters assigned to each atom follow from large scale quantum mechanical calculations. Note that there are additional variations of this paradigm in use such as

the “united-atom” model, which treats each terminal methyl group or intermediate methylene unit as a single particle, and the so-called “bead” model which assigns two to four particles to an individual amino acid in a protein system (Kuhn et al., 2000). It is the “all-atom” model however that finds use in this project via the OpenContact© software and, thus, the data that follows in the coming sections is entirely from an atomic perspective.

Now that the modeling technique has conceptually been defined, it is appropriate at this point to specifically detail the mathematical treatment of the “all-atom” model or potential function. From a mathematical and physical perspective, the molecular mechanics calculations are based simply a summation of a variety of different potentials inherent in a molecular system. The two most fundamental categories of these potentials are the collective potentials governing covalent bonding between atoms and the collective potentials governing non-covalent bonding (i.e. electrostatic attractions). These two master families (covalent and non-covalent) are then added together to give the total potential energy of the molecule under study. Therefore mathematically the functional form of the potential function is (Kollman, Massova, Reyes, & et al., 2000),

(1) $E_{TOT} = E_{COVALENT} + E_{NON-COVALENT}$

where the covalent term is,

(2) $E_{COVALENT} = E_{BOND} + E_{ANGLE} + E_{DIHEDRAL}$

and the non-covalent term is,

(3) $E_{NON-COVALENT} = E_{ELECTROSTATIC} + E_{VAN\ DER\ WAALS}$

Eqs. (2) and (3) contain some of the potentials briefly mentioned in the foregoing paragraphs and we now see where and how exactly these potentials are prescribed in the grand scheme of molecular mechanics calculations.

Each term in Eqs. (2) and (3), in turn, are defined according to specific potential models. Starting with Eq. (2) (modeling covalent bonding), the bond and angle terms (E_{BOND} and E_{ANGLE}) are typically modeled as harmonic potentials centered about equilibrium bond length values. This approach makes use of the classical setup commonly seen when analyzing the equations of motion for periodically occurring phenomena such as a swinging pendulum or (as is most appropriate in the present case of a chemical bond) a spring rapidly expanding and contracting. Such a harmonic potential may therefore be of the form,

(4) $x(t) = Acos(\omega t + \phi)$

where,

$$\omega = \sqrt{\frac{k}{m}} = \frac{2\pi}{T}$$

which would represent the angular frequency of bond vibration. The dihedral term in Equation 2 ($E_{DIHEDRAL}$) is modeled with the appropriate potentials, but they tend not to be harmonic oscillators and their various functional forms tend to vary with the specific implementation of the

potential function. Improper dihedral mathematical models may be included to enforce the planarity of aromatic rings and other conjugated systems.

For the non-covalent interactions considered in Eq. (3), the electrostatic term ($E_{ELECTROSTATIC}$) takes into account any inherent charges possessed by the constituent atoms of the protein due to electron deficiencies or excesses as well as dipole, quadrapole, or higher moments due to shifting electron positions. It is modeled with Coulomb's law for two point charges (Peters, 2007),

(5) $u(r) = \frac{q_1 q_2}{r}$

where q_1 and q_2 are the charges of each point and are given in electrostatic units (1 esu ≡ 1 dyne$^{1/2}$▪cm). The van der Waals term ($E_{VAN\ DER\ WAALS}$) in Equation 3 further considers BOTH the short range (Born) repulsive forces and the longer-range (van der Waals) attractive forces present at the intermolecular level (Peters, 2007). These forces, although similar in concept to the electrostatic forces previously mentioned, primarily differ in the sense that van der Waals forces are associated with oscillating, time-dependent electric fields created by the wave, or probabilistic behavior, of electrons in atoms (by themselves, van der Waals forces are mathematically modeled as r^{-7} where r is the separation distance between the two species) (Peters, 2007). This is in distinction to the more fixed nature of positive or negative charge that is modeled by Coulomb's law. Although the term in Eq. 3 primarily bears the name van der Waals, it is important to note that the specific potential function actually used incorporates both Born and van der Waals repulsions and attractions respectively. On the atomic scale, the precise nature of Born repulsive forces manifest themselves when two atoms (or molecules) approach

each other at distances slightly larger than σ where σ is the atomic or molecular diameter. At these very small separation distances, the electron shells of the two molecular (or atomic) species partially penetrate each other (Peters, 2007). The nuclei of the constituent atoms (which are positively charged) are now no longer shielded by their respective electron shells and so they now repel one another. It is important to note that the scenario of Born repulsion, despite the physical similarities with covalent bonding, differs from covalent bonding in that no electrons are shared by the two species in question in the Born set up. Both Born and van der Waals interactions are thus modeled in Eq.(3)'s $E_{VAN\ DER\ WAALS}$ term using the Lennard-Jones Potential (Stobener, Klein, Reiser, Horsch, Kufer, & Hasse, 2014),

(6) $$u(r) = 4\varepsilon\left[-\left(\frac{\sigma}{r}\right)^6 + \left(\frac{\sigma}{r}\right)^{12}\right]$$

where ε and σ are constants related to the energy minimum and distance at minimum, respectively.

3.2 FORCE FIELDS: MOLECULAR MECHANICS APPLIED

As stated in the previous section, the ultimate manifestation of the mathematical engine that drives OpenContact©'s calculations is the atomic force field and associated parameters. It is crucial to make the distinction however of what is meant by the term "force field" in chemistry and computational biology as opposed to its more common usage in physics. In chemistry, a force field refers to a system of potential energy functions like the kind that have been discussed in the Equations 2 and 3. This stands in contrast to the definition of a force field in physics as

being the gradient of a scalar potential (Peters, 2007). Another important concept to be cognizant of is that despite the fact that the equations in conjunction with their parameters are called "force fields", the quantities that the equations actually provide are potentials, not forces. Strictly speaking, the force is defined mathematically as the negative of the first derivative of the potential function with respect to distance. In addition to providing information on each element present in the molecule, the parameters also include specific constants for the properties of these atoms as they appear in different functional groups such as carbonyl and hydroxyl oxygens (Krall, Brunn, Kankanala, & Peters, 2014).

The parameter sets used in proper force field analysis include data for van der Waals radii, partial charges, atomic masses, equilibrium values for bond lengths and angles, dihedral angles for bonded atoms and the effective "spring constants" for each bond potential (recall that in molecular mechanics, bonds are well approximated by the physics of an actual expanding and contracting spring) (Krall et al., 2014). An important point of emphasis is that despite the fact that these values and constants are ultimately employed on a holistic perspective of the entire protein under study, the values themselves are derived primarily from quantum mechanical models on small and simple organic molecules (Cornell et al., 1995). As will be seen in the next section, there are many force field models currently available for intra and inter molecular interaction studies, and each model has its own set of parameters. These parameters are defined by force field developers to be self-consistent within the parameters' respective force field model and should therefore never be used in conjunction with the potential equations from another force field model, no matter how similar parts of the models may be.

3.3 AMBER 03: OPENCONTACT© DEFINED

The specific force field model used by OpenContact© is the AMBER 03 force field. AMBER is an acronym standing for **A**ssisted **M**odel **B**uilding with **E**nergy **R**efinement. AMBER 03 follows the skeletons of Eqs. (2) and (3) and it has its own set of parameters for van der Waals radii, bond lengths, etc…. There do exist other members of the AMBER family of force fields that are popularly used but as mentioned earlier, each has its own exclusive parameter sets that will only work with the force field model to which it belongs and it is this proprietary attribute that defines the differences between individual members of the AMBER family (Cornell et al., 1995).

Given that the parameter sets are the only attributes that differ between AMBER family members, it should come as no surprise then that the foundational molecular mechanics potential function for the AMBER family of force fields remains the same for each currently available AMBER family member. The actual functional form of the AMBER equation is (Cornell et al., 1995),

$$(7) \quad V(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2$$

$$+ \sum_{\text{torsions}} \sum_{n} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)]$$

$$+ \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \varepsilon_{ij} \left[\left(\frac{\sigma_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right\}$$

For the right hand side of this equation, the first term, which sums over bonds, models the potential energy between covalently bonded atoms (Cornell et al., 1995). The classical spring model is a good approximation for atoms participating in bonds, so long as the bound atom distance (l) remains relatively close to the equilibrium bond length (l_0). This model begins to give poorer results, however, as the distance of binding is increased (Cornell et al., 1995). The second term, which sums over the angles of electron orbitals (θ), models the potential energy of electron orbitals involved in covalent bonding (Cornell et al., 1995). The third term, which sums over torsion angles (ω), models bond rotation as influenced by bond order (i.e. double bonds vs. single bonds) and neighboring charge clouds (i.e. vicinal bonds on the same atom and electron lone pairs) (Cornell et al., 1995). It is possible for any given individual bond to have more than one of these terms and in such cases, the total torsional energy potential would be expressed as a Fourier series (Cornell et al., 1995). The fourth and last term (the double summation over *i* and *j*) models the interactions arising from non-covalent force sources where r_{ij} is the distance between the two interacting atoms in the system (Peters, 2007). Recall that these forces were identified as being modeled by the Lennard-Jones potential for those non-covalent interactions

arising from time-dependent electron oscillation phenomena and Coulomb's law for charged atomic species.

4.0 MATERIALS AND METHODS: PRIMARY CONSIDERATIONS

At its foundation, this research project was purely computational and entailed feeding raw protein structure data files into a suite of software applications to provide the desired output data. This data then had to be compared (wild-type against mutant) to identify any differences between wild-type and mutant strains. Ultimately, this required understanding the systems that were going to be under study and establishing test subjects (the input data files) that had minimal variability from one to the other within its respective system (a system being defined as a wild-type strain in conjunction with its mutant counterpart). Having an availability of both wild-type and mutant data files in the first place was another criterion that had to be considered and satisfied before a system could be approved for study.

Minimizing variability was a very important consideration in identifying which protein systems would be analyzed. Each system had to, first and foremost, have its wild-type and mutant strains come from the same organism. There were a plethora of promising systems that either came from cattle, mice or even invertebrates such as insects and worms, but since the focus of this research project was disease in humans, the ultimate candidates chosen had to come from *Homo sapiens*, the species of man. The next criterion considered in minimizing variability within each system was data file similarity. As shall be discussed shortly in greater detail, the input data files essentially map the positions of each constituent atom of their respective proteins in space. In many cases, there were multiple input structure data files for the same protein, from

the same species, but which covered/mapped entirely different regions of said protein. Naturally, a system comprised of dissimilar locations for the wild-type and mutant strains would be of no use to a systematic and rationalized study and so input structure data files had to be found that had as much overlap as possible. An approved system would thus finally consist of both a wild-type and mutant strain (thus satisfying availability), and a wild-type and mutant set of structure data files that had as much protein region overlap as possible (thus satisfying variability minimization). Finally, it was decided that any system considered should not contain a cysteine amino acid mutation, since these mutations can drastically alter the three-dimensional structure of the protein, and the focus of this study was more on the subtle changes that could be incurred from amino acid replacements that did not necessarily change protein structure. Although residue mutations are the featured aberrancy in this study, it is important to keep in mind that other factors associated with the environment of the system (i.e. pH, presence of solutes, temperature etc…) can also have a profound effect on protein behavior. Thus this research project only looked at just one aspect of how mutation states can affect protein-protein interactions.

4.1 MATERIALS AND METHODS: PDB FILE ANATOMY

The input data files came from the Protein Databank or pdb website (www.rcsb.org) and are consequently called "pdb files." They are logged and registered in the databank according to an accession number unique to each protein entry. These accession numbers are usually a combination of alpha-numeric characters and they constitute the file name upon downloading. The pdb files are inherently text files that can be viewed with any simple text editor. Notepad++ was specifically used to view and study the pdb files due to this piece of software's designed

purpose of viewing machine instructions in interpreted code (i.e. high-level programming languages). Although fundamentally pdb files bear the ".txt" file extension, they have an additional appellative of ".pdb" following the normative ".txt" extension. It is the ".pdb" sub-extension that allowed these input data files to be read by the various computer programs used which will be discussed in more detail in the coming sections. Here is a sample pdb file name as would be downloaded from the Protein Databank: 4BXO.pdb.

What type of information specifically is encoded in these pdb files? The answer to this question is a set of coordinates that detail where in space (a Cartesian-like x-y-z volume) each and every constituent atom of the protein is located. This information is gathered from X-ray crystallography studies, NMR assays or neutron diffraction studies. None of the test cases reported here had their structures elucidated from the neutron diffraction assay however. From the visualization software's (DS Viewer Pro's) perspective, this information is plotted and each successive atom or point is connected to yield the overall tertiary structure of the protein. For multi-chain proteins (which have each separate chain encoded in the same pdb file), each individual chain's atoms are also plotted and connected in the same space and the resulting quaternary structure is then presented to the user. The section of the pdb file that encodes atom positions obeys the skeleton as shown in Fig. A1 (see appendix). Below is the definition for each column entry:

- Column1: field for the actual species to be plotted, which for this example is an atom.
- Column 2: field for the sequential number of the corresponding atom (the value of this field will naturally increase by one as one moves down Column 2).

- Column 3: field for the specific atom name with respect to the amino acid residue to which the atom belongs (i.e. the alpha carbon, amino nitrogen, carbonyl oxygen, etc…).
- Column 4: field for the standard three-letter amino acid code for the residue.
- Column 5: field for the protein chain that particular residue's atom belongs to.
- Columns 7, 8 & 9: fields for the x-y-z coordinates respectively of each atom.
- Column 10: field for the occupancy value of the atom.
- Column 11: field for the thermal factor of the atom.
- Column 12: field for the explicit chemical element of the atom (i.e. carbon, nitrogen etc…).

This example was taken from the transthyretin protein (accession number: 3D7P), one of the test cases. Following the plotting of the above data, DS Viewer Pro then allows rotation of the image to see the three dimensional structure in depth and from different angles. There are a host of other features likewise available in DS Viewer Pro that allow for selective hi-lighting of amino acid residues, labeling the distinct chains (if there are more than one) and zooming in to regions of interest on the protein, just to name a few. It is important to keep in mind however that despite the fact that the pdb file encodes raw atom positions, the output labelling is primarily from the perspective of the amino acid residue.

In addition to atom positions, the typical pdb file will also include information on the type of experiment that yielded the protein's structural results (i.e. X-ray diffraction, NMR, neutron scattering), the resolution of the imaging, the position of mutations in the primary structure of each protein chain (if the pdb file is of a mutant strain), the name of the protein, any missing atoms or residues, and the names of the authors who elucidated the structure.

Depending on the depth of study from the researchers who compiled the pdb file, there may even be a section detailing additional suggested pdb accession numbers for related protein structures. Figures A1 and A2 show snapshots of the two important sections of a typical pdb file (taken again from transthyretin-3D7P),

4.2 MATERIALS AND METHODS: OVERALL EQUIPMENT

Due to the computational nature of this research project, the principal pieces of equipment consisted of two main classes: computer hardware and software. Hardware-wise, a Dell® Inspiron 15 laptop computer with an Intel® Inside™ CORE™ i3 processor with 4 gigabytes of RAM was used. Data was saved on said machine as well as on various disk drives for backup in case of primary storage failure. Software-wise, OpenContact© was used to process raw input data (the pdb structure files from the Protein Databank) and consequently provided quantitative output calculation results for criteria such as inter-atomic distances and Lennard-Jones Potentials which was then processed and organized using Micrsoft Excel spreadsheeting. OpenContact© also provided qualitative contact mapping plots for the same inter-atomic criteria mentioned above as well as molecular visualization files of the output data for qualitative analysis as well. Consequently, the other indispensable piece of software used was DS Viewer Pro®, a molecular visualization program that would read the same input files that were fed to OpenContact©, but instead provided output in the form of a 3 dimensional image of the protein encoded by the input data. As with many available molecular visualization software, DS Viewer Pro® permits for the rotation of the three dimensional images so as to allow viewing the molecules from different angles.

4.3 MATERIALS AND METHODS: DS VIEWER PRO®-THE FIRST STEP

Once a protein system was selected, the first step in analysis was looking at the entire structure of both wild-type and mutant strains. Such a qualitative approach served to primarily gain an appreciation and knowledge of the location of the general residues potentially involved in intra-protein binding. The software that was used to provide these molecular visualizations was DS Viewer Pro by Dassault Systems BioVia. Despite the fact that there were a number of other molecular visualization programs available for potential use, DS Viewer Pro was ultimately selected due to the publication quality of its output data files. The actual artistic representation style of the protein structure deployed by DS Viewer Pro was in the hands of the user. The most common method seen thus far by the author in text-books and scientific articles is the flat ribbon representation. Below is a sample of this representation of molecular structure, again, for transthyretin, accession number 3D7P,

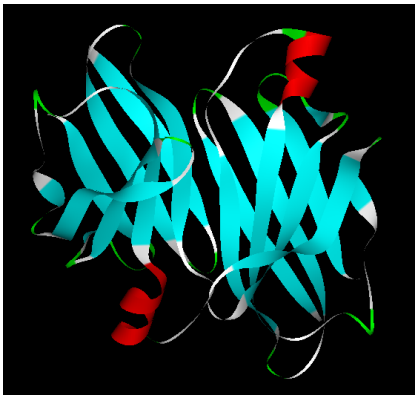


Fig. 1: Flat ribbon representation of transthyretin (wild-type).

As alluded to earlier, DS Viewer Pro Due provides many features to the user in displaying protein structure from pdb files. One such feature that proved to be most curious but useful in

studying the detailed position of residues and their constituent atoms was the option of overlaying a wire-frame representation of each and every atom present in the pdb file over the general flat ribbon structure. The same transthyretin structure with its atoms overlaid the ribbon representation is shown in Figure A3. Other representation schema available included wire, stick, line ribbon and schematic, to name a few. The same transthyretin protein is shown in Figs. A4-A7 in these different representations. It should be noted however that in some specific instances to be seen later, the schematic representation (Fig. A7) proved to be more appropriate.

One of the more powerful and informative features of DS Viewer Pro was the ability to see and selectively hi-light residues in a protein structure. The following image showcases this ability for the same wild-type transthyretin model with the specific amino acid that was to be mutated in the mutant pdb file, here hi-lighted yellow,

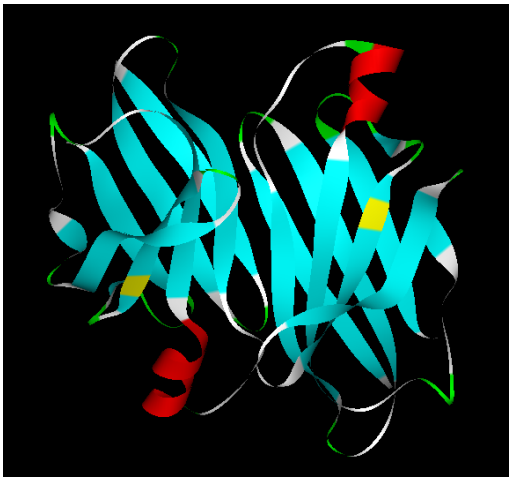


Fig. 2: Transthyretin with valine at the 30th position hi-lighted

Next is the same image, but with the residue names showing. Again, the same amino acid is hi-lighted in yellow.

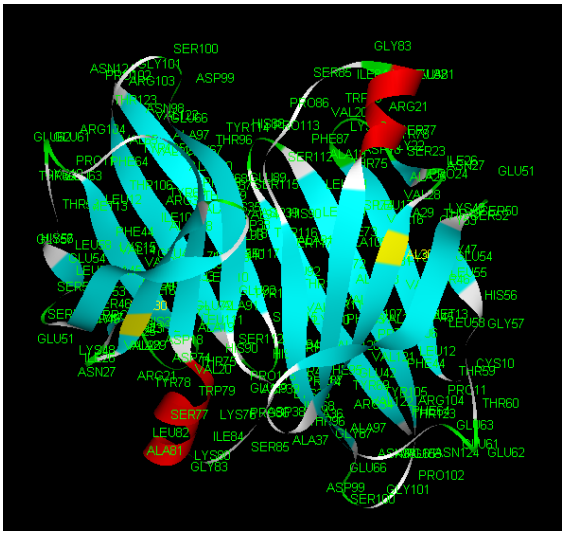


Fig. 3: Transthyretin with all residues displayed; valine at 30th position hi-lighted yellow

It is important to note however that select residues such as valine at the 30th position cannot be solely displayed all the while ignoring (i.e. not displaying) the other residues in the protein. If such images are desired, a manual labeling via the use of superimposed text boxes over the bare protein structure would have to be employed.

4.4 MATERIALS AND METHODS: OPENCONTACT©- THE NEXT STEP

Following the preliminary structure viewing of the overall protein system, the next step was to execute OpenContact© on both the wild-type and mutant strains for the given system under study. OpenContact© is a simple, yet powerful static contact mapping algorithm that has been developed as a first step at identifying binding "hot-spots" in protein-protein interactions from interaction partner structure files (Krall et al., 2014). The algorithm is fast and provides results on a Windows© based laptop or PC within approximately two minutes even for large protein systems. It has been demonstrated that OpenContact© is capable of identifying key or

critical inter-atomic hot-spots in PPI's previously obtained via more sophisticated methods such as MD or docking methods (Krall et al., 2014). Again, such results furnished by this algorithm were provided very quickly within a two minute time window and it additionally had the benefit of not requiring the user to have any sophisticated knowledge associated with carrying out large scale MD or docking methods. Thus, a large number of systems can be quickly mapped as an initial step to potentially ascertain the roles of PPI's in protein behavior. What now follows is a general overview of the input protocol and considerations for inputting the pdb data file into OpenContact©. As has been discussed thus far, OpenContact© calculates inter-atomic distances, and force potentials by reading in the pdb files containing atomic position data. The exact nature by which this is done is by first specifying the two chains of the protein that one wishes to study (only two chains can be analyzed at any given algorithm execution however). Below is a snapshot of the graphical user interface (GUI) of OpenContact© by which this first step is taken,

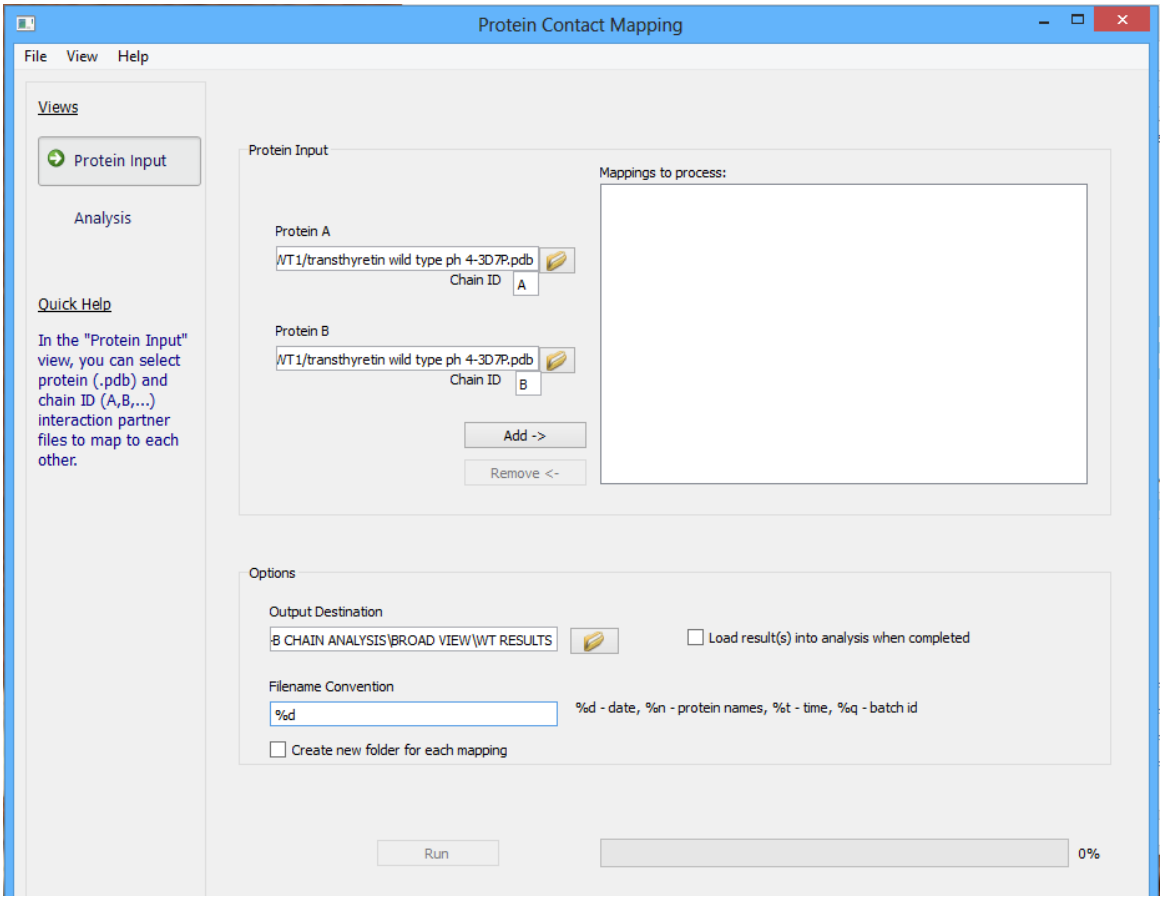


Fig. 3: OpenContact© GUI

Several details worth noting from this image are the two populated browse input fields located in the section marked “Protein Input.” It is in this field that the user would specify (most conveniently by using the browse buttons with folder icons) the directory pathway of the input pdb files. Again, there are two pathway fields to allow the study of the interactions of two chains (Krall et al., 2014). Since the backbone of these computational calculations is the actual structure encoded by the pdb file, the two pathways will naturally direct to the exact same file. This will not always be the case however, especially with single chain protein systems as will be seen shortly. The next item to consider in Fig. 3, still in the “Protein Input” section, is the “Chain ID” input text box located immediately underneath each directory pathway. As this field’s name

suggests, the appropriate input to this field is the specific chain one wishes to include in the computation study (Krall et al., 2014). It is absolutely important that the chains specified are actually logged in the pdb file however, as the program will be unable to proceed if it is unable to find what is provided. Once directory pathways and chain ID's have been specified, the "Add" button must then be clicked to formally feed these results into OpenContact©. Following pdb file and chain specifications, the next two items of consideration are now in the section titled "Options." The pathway field marked "Output Destination" would point to a directory location where the user would like to store the output files generated by OpenContact©. The text box marked "Filename Convention" would direct OpenContact© with the paradigm for naming the output files (Krall et al., 2014). As can be seen, file-naming by date of creation, protein names and batch id are some options left to the user to choose from. Finally, the radio button marked "Create new folder for each mapping" does as its name suggests: if the same overall destination folder is maintained for separate runs, each run will be stored in a newly created folder automatically generated by OpenContact© (Krall et al., 2014). Once each of the aforementioned fields have been populated and the "Add" button depressed, the last and final step of the data input protocol is to depress the bottom-most button marked "Run." A green bar will then qualitatively indicate the stage of the program's data processing (i.e. the green bar will grow as the process approaches completion) and once the procedure is complete, the output files will be stored in the directory specified.

The precise nature of the output data files is essentially divided along two criteria: fine parsing and coarse parsing, of the calculated potentials and distances (Krall et al., 2014). For the coarse parsing criterion, atom-atom separation distances greater than 10.5 angstroms are excluded from the final data file (Krall et al., 2014). In the fine parsing of the interactions, an

additional restriction on the atom-atom interaction potentials are specified. For all the results in the coming sections, the upper limits to restrict Coulombic and Lennard-Jones interactions are (respectively) (Krall et al., 2014),

(8) $U^*_{Coul} \leq -0.3$

or,

(9) $U^*_{LJ} \leq -0.1$

where,

(10) $U^* = \frac{U}{kT}$

for both interaction criteria. In Equation 10, *k* is Boltzmann's constant and *T* is temperature taken as 310.15 K for all results that follow (Krall et al., 2014). Consequently, the values reported by OpenContact are dimensionless where potentials have been scaled by *k*T and distance has been scaled by 1 nano-meter. These potential energy cut-offs ensure that only the strongest attractive interaction potentials for any given type are included in the fine parsing results which are provided to the user in a highly manageable and user-friendly format in both ".pdb" and ".txt" file types (the ".txt" file is purely a text file that allows its data content to be properly imported into Microsoft Excel for spreadsheeting) (Krall et al., 2014). The data can now be analyzed, plotted and generally manipulated according to the user's needs as shall be

seen in the coming cases. At its most basic, the fine parsing data thus provides which atom-atom interaction pairs exhibit the strongest interactions, whereas the coarse parsing data provides which atom-atom interaction pairs contribute most to the structure involved in the inter-chain bound region (Krall et al., 2014). For the studies that follow in this research project, the coarse parsing data provided the greatest insight into the structural changes exhibited by mutant strains compared to their wild-type counterparts.

Following the above data input protocol, the user is now able to do one of two things: view and manipulate the output data files, or create a contact mapping plot of the two chains based on Coulombic, Lennard-Jones or distance criteria. This plot inherently uses the output data from the fine parsing files already generated but it presents that data in a pictorial format that provides a very nice overview of the atom-atom interactions between the two chains specified. Below is a snapshot of such a plot, again, for wild-type transthyretin, accession number 3D7P,

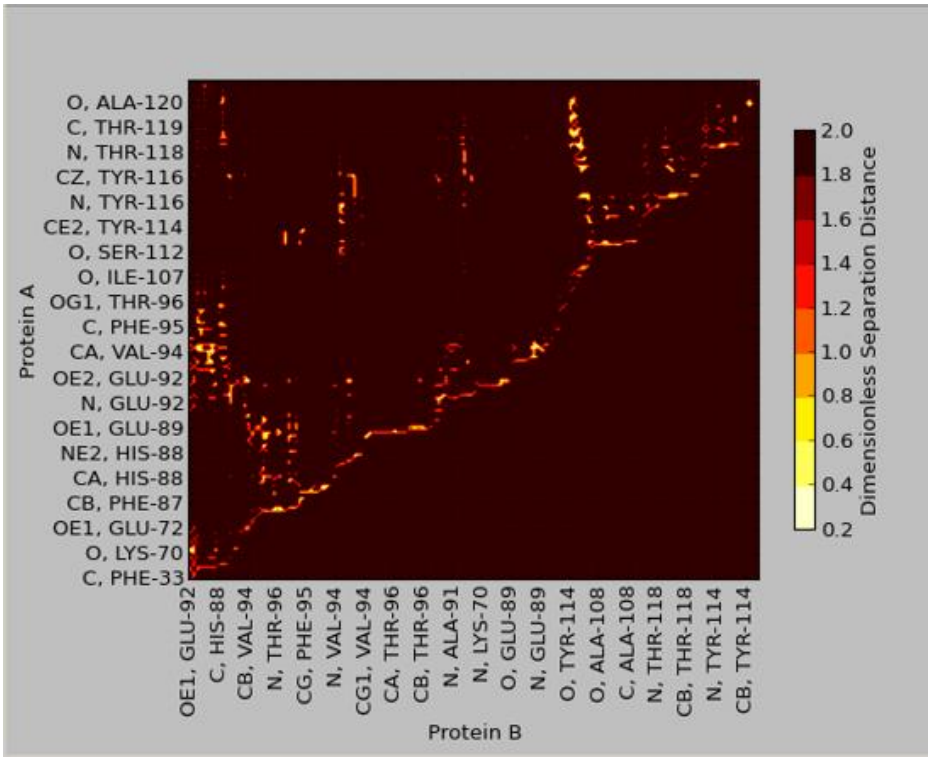


Fig. 4: Contact map plot for distances between atoms in chains A and B for transthyretin.

As can be seen in Fig. 4, the corresponding x and y axes display specific atoms involved in the interactions between chains A and B for transthyretin. The y axis (marked "Protein A") presents the participant atoms from the A chain in ascending order (from the bottom of the axis to the top of the axis) and the x axis (marked "Protein B") presents the corresponding atomic partners from the B chain that interact with the A chain atoms. The atoms in the x axis are not in any particular order however. It is important to also note that despite the fact that a certain number of atoms may be explicitly presented in the above plot, the actual fine parsing analysis may detect hundreds more. Again, these plots are more qualitative in nature, but they can provide a very good overall picture of how the system's wild-type or mutant strains interactions behave. The informative nature of the plot ultimately comes from the bright colors in the actual plot area:

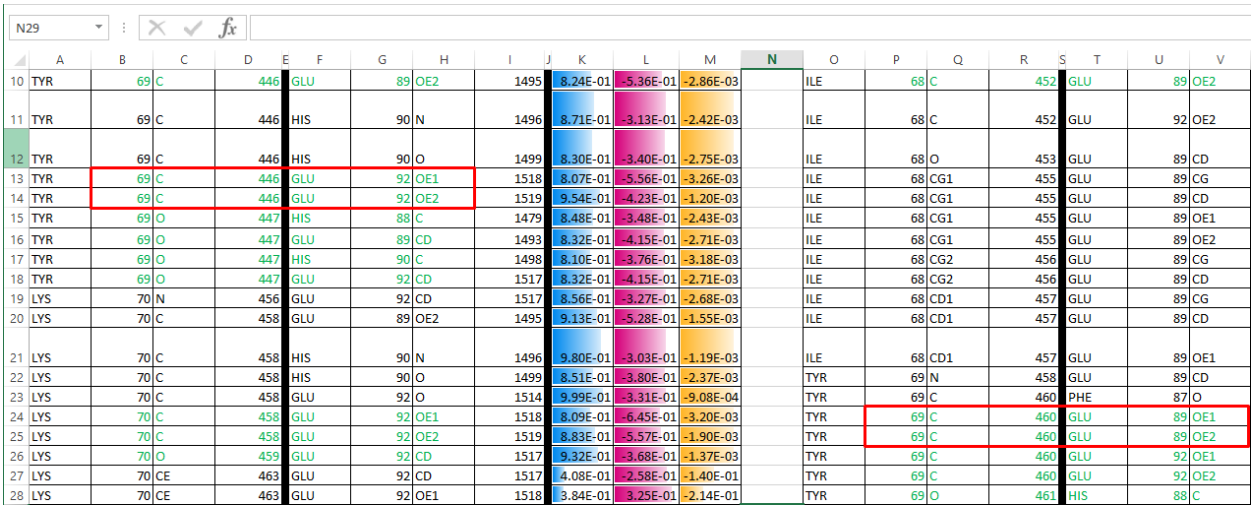
brighter colors at certain ordinate and abscissa intersections indicate stronger interactions, or, as is the case for Fig. 4, closer separation distances. The color-coded legend to the right of the plot in the figure provides a neighborhood for the corresponding values. Although powerful, for the present application, these contact plots were not used, as structure image files were the central starting point for analysis of the input files and results.

As briefly mentioned in the preceding paragraphs, OpenContact's cut-off limits were determined to be efficient for the research project undertaken by Peters et al. which was finding regions of certain proteins involved in disease that were amenable to synthetic peptide (bio-mimetic peptide) targeting for therapeutic use. A total of six unrelated protein systems were analyzed in the Krall et al. study and included gp41 (a glycoprotein belonging to HIV that is necessary for host cell entry), the EpherinB4-EpherinB2 system (a cell-cell signaling angiogenesis stimulation mechanism), Myc-Max (a transcription factor hetero-dimer complex involved in cell proliferation), and EGFR-EGF (a cell growth factor system). According to the results of the study, the OpenContact© algorithm "efficiently predicted proposed peptide biomimetics, or close variations thereof" when compared against peptide mimetics that had been proposed and/or developed for these systems to some degree of success by purely experimental methods (Krall et al., 2014).

4.5 MATERIALS AND METHODS: OUTPUT FILE ANATOMY

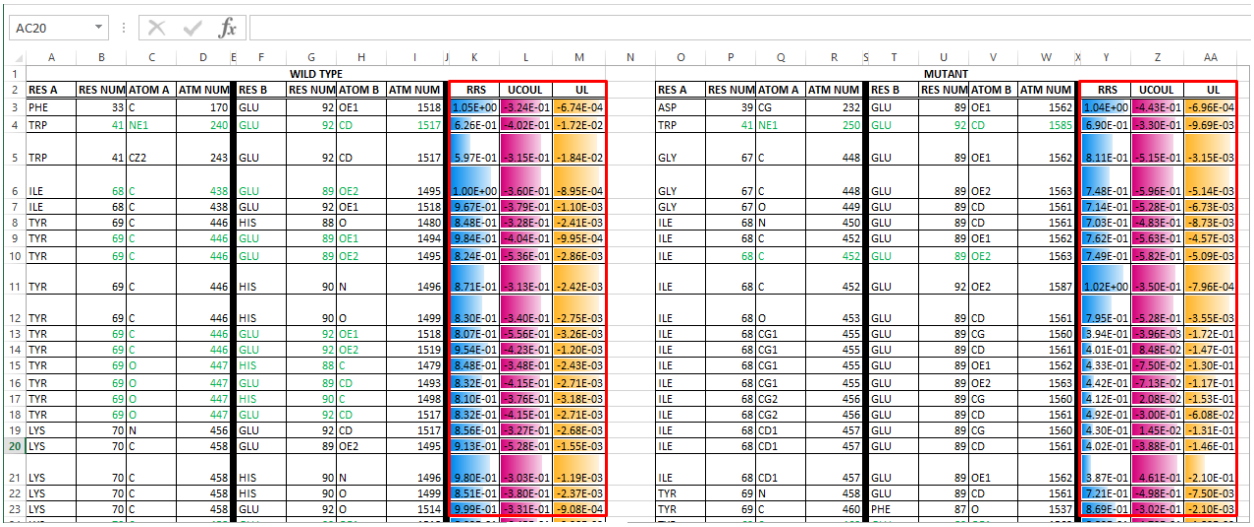
Since the primary focus of this study was interaction changes from the wild-type to the mutant version of the protein, the discussion that follows comes from the results of the fine

parsing output data. Coarse parsing data obeys the same format however. These results generated by OpenContact© were consequently interpreted in the following manner. First the entire protein system was analyzed holistically by summing the two potential criteria (Coulombic and Lennard-Jones) for both the wild-type and mutant strains. This permitted seeing how the totality of interactions changed on a big picture scale. The interatomic distances by contrast were averaged to get a sense of how average distances changed. Then, common interaction partners amongst the two strains were identified and only the common interaction potentials and distances were summed and averaged respectively as described for the holistic scenario. From the fine parsing data (imported into Microsoft Excel), common interaction partners were hi-lighted in green to differentiate them from interactions unique to each strain. Two corresponding common entries for transthyretin's fine parsing analysis are shown in Figs. 5 and 6.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
10	TYR	69	C	446	GLU		89	OE2	1495		8.24E-01	-5.36E-01	-2.86E-03		ILE	68	C	452	GLU		89	OE2
11	TYR	69	C	446	HIS		90	N	1496		8.71E-01	-3.13E-01	-2.42E-03		ILE	68	C	452	GLU		92	OE2
12	TYR	69	C	446	HIS		90	O	1499		8.30E-01	-3.40E-01	-2.75E-03		ILE	68	O	453	GLU		89	CD
13	TYR	69	C	446	GLU		92	OE1	1518		8.07E-01	-5.56E-01	-3.26E-03		ILE	68	CG1	455	GLU		89	CG
14	TYR	69	C	446	GLU		92	OE2	1519		9.54E-01	-4.23E-01	-1.20E-03		ILE	68	CG1	455	GLU		89	CD
15	TYR	69	O	447	HIS		88	C	1479		8.48E-01	-3.48E-01	-2.43E-03		ILE	68	CG1	455	GLU		89	OE1
16	TYR	69	O	447	GLU		89	CD	1493		8.32E-01	-4.15E-01	-2.71E-03		ILE	68	CG1	455	GLU		89	OE2
17	TYR	69	O	447	HIS		90	C	1498		8.10E-01	-3.76E-01	-3.18E-03		ILE	68	CG2	456	GLU		89	CG
18	TYR	69	O	447	GLU		92	CD	1517		8.32E-01	-4.15E-01	-2.71E-03		ILE	68	CG2	456	GLU		89	CD
19	LYS	70	N	456	GLU		92	CD	1517		8.56E-01	-3.27E-01	-2.68E-03		ILE	68	CD1	457	GLU		89	CG
20	LYS	70	C	458	GLU		89	OE2	1495		9.13E-01	-5.28E-01	-1.55E-03		ILE	68	CD1	457	GLU		89	CD
21	LYS	70	C	458	HIS		90	N	1496		9.80E-01	-3.03E-01	-1.19E-03		ILE	68	CD1	457	GLU		89	OE1
22	LYS	70	C	458	HIS		90	O	1499		8.51E-01	-3.80E-01	-2.37E-03		TYR	69	N	458	GLU		89	CD
23	LYS	70	C	458	GLU		92	O	1514		9.99E-01	-3.31E-01	-9.08E-04		TYR	69	C	460	PHE		87	O
24	LYS	70	C	458	GLU		92	OE1	1518		8.09E-01	-6.45E-01	-3.20E-03		TYR	69	C	460	GLU		89	OE1
25	LYS	70	C	458	GLU		92	OE2	1519		8.83E-01	-5.57E-01	-1.90E-03		TYR	69	C	460	GLU		89	OE2
26	LYS	70	O	459	GLU		92	CD	1517		9.32E-01	-3.68E-01	-1.37E-03		TYR	69	C	460	GLU		92	OE1
27	LYS	70	CE	463	GLU		92	CD	1517		4.08E-01	-2.58E-01	-1.40E-01		TYR	69	C	460	GLU		92	OE2
28	LYS	70	CE	463	GLU		92	OE1	1518		3.84E-01	3.25E-01	-2.14E-01		TYR	69	O	461	HIS		88	C

Fig. 5: Fine parsing analysis data for transthyretin. The wild type results are in the left-hand table, mutant results are at the right. Two sample common interaction entries are bordered in red.



	WILD TYPE											MUTANT									
	RES A	RES NUM	ATOM A	ATM NUM	RES B	RES NUM	ATOM B	ATM NUM	RRS	UCOUL	UL	RES A	RES NUM	ATOM A	ATM NUM	RES B	RES NUM	ATOM B	ATM NUM	RRS	UCOUL	UL
3	PHE	33	C	170	GLU	92	OE1	1518	-1.05E+00	-3.24E-01	-6.74E-04	ASP	39	CG	232	GLU	89	OE1	1562	-1.04E+00	-4.43E-01	-6.96E-04
4	TRP	41	NE1	240	GLU	92	CD	1517	-6.26E-01	-4.02E-01	-1.72E-02	TRP	41	NE1	250	GLU	92	CD	1585	-6.90E-01	-3.30E-01	-9.69E-03
5	TRP	41	CZ2	243	GLU	92	CD	1517	-5.97E-01	-3.15E-01	-1.84E-02	GLY	67	C	448	GLU	89	OE1	1562	-8.11E-01	-5.15E-01	-3.15E-03
6	ILE	68	C	438	GLU	89	OE2	1495	-1.00E+00	-3.60E-01	-8.95E-04	GLY	67	C	448	GLU	89	OE2	1563	-7.48E-01	-5.96E-01	-5.14E-03
7	ILE	68	C	438	GLU	92	OE1	1518	-9.67E-01	-3.79E-01	-1.10E-03	GLY	67	O	449	GLU	89	CD	1561	-7.14E-01	-5.28E-01	-6.73E-03
8	TYR	69	C	446	HIS	88	O	1480	-8.48E-01	-3.28E-01	-2.41E-03	ILE	68	N	450	GLU	89	CD	1561	-7.03E-01	-4.83E-01	-8.73E-03
9	TYR	69	C	446	GLU	89	OE1	1494	-9.84E-01	-4.04E-01	-9.95E-04	ILE	68	C	452	GLU	89	OE1	1562	-7.62E-01	-5.63E-01	-4.57E-03
10	TYR	69	C	446	GLU	89	OE2	1495	-8.24E-01	-5.36E-01	-2.86E-03	ILE	68	C	452	GLU	89	OE2	1563	-7.49E-01	-5.82E-01	-5.09E-03
11	TYR	69	C	446	HIS	90	N	1496	-8.71E-01	-3.13E-01	-2.42E-03	ILE	68	C	452	GLU	92	OE2	1587	-1.02E+00	-3.50E-01	-7.96E-04
12	TYR	69	C	446	HIS	90	O	1499	-8.30E-01	-3.40E-01	-2.75E-03	ILE	68	O	453	GLU	89	CD	1561	-7.95E-01	-5.28E-01	-3.55E-03
13	TYR	69	C	446	GLU	92	OE1	1518	-8.07E-01	-3.56E-01	-3.26E-03	ILE	68	CG1	455	GLU	89	CG	1560	-8.94E-01	-3.96E-03	-1.72E-01
14	TYR	69	C	446	GLU	92	OE2	1519	-9.54E-01	-4.23E-01	-1.20E-03	ILE	68	CG1	455	GLU	89	CD	1561	-8.01E-01	-8.48E-02	-1.47E-01
15	TYR	69	O	447	HIS	88	C	1479	-8.48E-01	-3.48E-01	-2.43E-03	ILE	68	CG1	455	GLU	89	OE1	1562	-8.33E-01	-7.50E-02	-1.30E-01
16	TYR	69	O	447	GLU	89	CD	1493	-8.32E-01	-4.15E-01	-2.71E-03	ILE	68	CG1	455	GLU	89	OE2	1563	-8.42E-01	-7.13E-02	-1.17E-01
17	TYR	69	O	447	HIS	90	C	1498	-8.10E-01	-3.76E-01	-3.18E-03	ILE	68	CG2	456	GLU	89	CG	1560	-8.12E-01	-2.08E-02	-1.53E-01
18	TYR	69	O	447	GLU	92	CD	1517	-8.32E-01	-4.15E-01	-2.71E-03	ILE	68	CG2	456	GLU	89	CD	1561	-8.92E-01	-3.00E-01	-6.08E-02
19	LYS	70	N	456	GLU	92	CD	1517	-8.56E-01	-4.27E-01	-2.68E-03	ILE	68	CD1	457	GLU	89	CG	1560	-8.30E-01	-1.45E-02	-1.31E-01
20	LYS	70	C	458	GLU	89	OE2	1495	-9.13E-01	-5.28E-01	-1.55E-03	ILE	68	CD1	457	GLU	89	CD	1561	-8.02E-01	-3.88E-01	-1.46E-01
21	LYS	70	C	458	HIS	90	N	1496	-9.80E-01	-3.03E-01	-1.19E-03	ILE	68	CD1	457	GLU	89	OE1	1562	-8.87E-01	-4.61E-01	-2.10E-01
22	LYS	70	C	458	HIS	90	O	1499	-8.51E-01	-3.80E-01	-2.37E-03	TYR	69	N	458	GLU	89	CD	1561	-7.21E-01	-4.98E-01	-7.50E-03
23	LYS	70	C	458	GLU	92	O	1514	-9.99E-01	-3.31E-01	-9.08E-04	TYR	69	C	460	PHE	87	O	1537	-8.69E-01	-3.02E-01	-2.10E-03

Fig. 6: Fine parsing data hi-lighting actual data content. The columns headed by "RRS" (blue), "UCOUL" (magenta) and "UL" (orange) contain the data for distance, Coulombic potentials and Lennard-Jones potentials respectively.

What is ultimately meant by common interaction partners refers solely to the entries headed by "RES A," "RES NUM," "ATOM A," and "RES B," "RES NUM," "ATOM B" in the column titles in Figs. 5 and 6. Should the individual entries in these columns within one table identically coincide with entries in the same columns in the other table, then those elements within the two tables were considered common interaction pairs. As mentioned earlier, some structural elucidation assays may result in certain atoms or residues not being detected and thus not reconciled in the pdb structure file. This would result in those residues or atoms being omitted from the orderly enumeration dictated by the "ATM NUM" column. It is for this reason that this column is not used in determining common interaction partners: individual atom numbering may change from experiment to experiment, but residue numbering never does.

4.6 MATERIALS AND METHODS: DATA RECONCILIATION

Since the ultimate goal of this research project was to compare the interactions between wild-type and mutant strains of the same protein, the overall approach was to sum the three interaction criteria together (or average them for the distance criterion) and compare the results. In turn, these interaction totals were carried out over three analysis considerations: an overall comparison, a common atom comparison and a non-common comparison. The overall comparison was simply a total of each of the interaction criteria (Lennard-Jones potential, Coulombic potential and distance) and the total of each criterion was compared between the wild-type and mutant strains. The common atom comparison consisted of identifying the atom-atom interactions that were found in both the wild-type and mutant strains (green text in Figures 5 and 6), summing these values for each strain and then comparing the sums. The non-common comparison in turn consisted of summing all atom-atom interactions that were unique to each strain, and likewise comparing these sums.

Once these sums were executed, the comparison of the data had to have some threshold for what would be considered significant and what could be disregarded as deviations arising from experimental artifacts (these artifacts coming from the conditions employed in the crystallography experiments). Normal artifacts due to experimental conditions that can vary on the atomic scale when crystallographic procedures are being executed is typically around 5-7%. Thus, the threshold for consideration was decided to be a minimum of approximately 10% difference between comparable criteria of both wild-type and mutant strains.

5.0 SYSTEMS ANALYZED: TRANSTHYRETIN

The first system considered in the analysis was the protein transthyretin. Transthyretin in its physiological state is a tetramer made up of two sub-units where each subunit is in turn a homo-dimer. The molecular weight of each monomer is 15.89 kDa for a total molecular weight of 63.56 kDa for the biologically active tetramer ("TTR(human)," 2014). For this study, the homo-dimer was selected according to the availability, non-cysteine residue mutation and minimization of variability criteria discussed earlier. Below is an image for the wild-type and mutant strains studied, accession numbers 3D7P and 3DJT respectively,

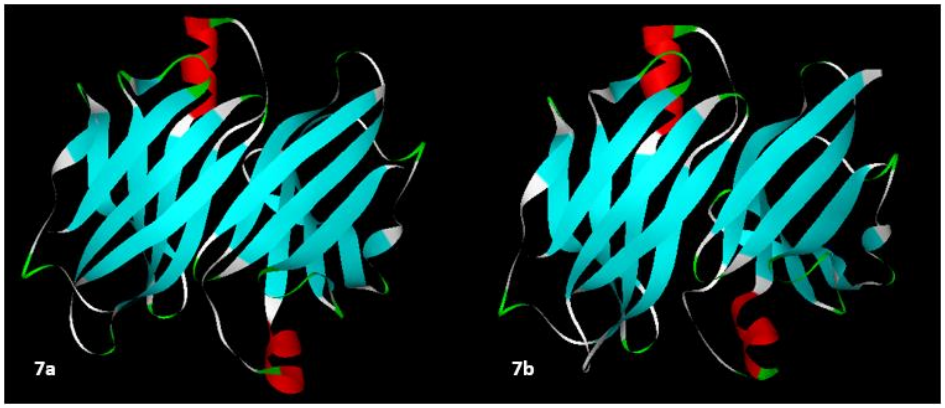


Fig. 7: Dimer of transthyretin wild-type (7a) and mutant (7b). Accession numbers 3D7P and 3DJT respectively.

Figures A10 and A11 show the biologically active transthyretin tetramer as well as the above dimer from a different angle.

Transthyretin is a carrier protein of two species: the thyroid hormone thyroxin and the lipocalin retinol binding protein which is itself bound to retinol (a form of vitamin A). It is from the transport of these two molecules that transthyretin derives its name: **trans**ports **thy**roxin and

retinol (Palinathan, Mohamedmohaideen, Snee, Kelly, & Sacchettini, 2008). The liver secretes transthyretin into the blood and the choroid plexus transports it into the cerebrospinal fluid. Thyroxin is a tyrosine-based hormone that acts on almost every cell in the body and ultimately plays a role in regulating the basal metabolic rate, long bone growth, neural maturation and even protein synthesis (Palinathan et al., 2008). It is produced in the thyroid gland and requires iodine for its synthesis (Palinathan et al., 2008). Incidentally, an iodine deficiency results in under-production of this hormone which in turn results in enlarged thyroid glands, the condition of which is known as simple goiters. One of the biological significances of transthyretin therefore lies in its ability to distribute thyroxin throughout the body so that the aforementioned biological processes of metabolism and cellular growth rate may not be impaired (Palinathan et al., 2008). The second chemical species transported by transthyretin, retinol, is responsible for a variety of functions such as maintaining vision, skin health, teeth re-mineralization, and bone growth (Palinathan et al., 2008). The exact avenue by which retinol affects these diverse systems depends on what form retinol (an alcohol) is converted to. As an aldehyde (retinal) it affects vision, as a carboxylic acid (retinoic acid) it affects skin health, teeth re-mineralization and bone growth (Palinathan et al., 2008).

Of important consequence to this study is the location of the active site for this and all the proteins that follow. In total, transthyretin carries one thyroxin molecule, and two retinol molecules via the retinol binding protein. For transthyretin, the studies by Monaco showed the active sites for the biologically active tetramer to be located on the outside of the complete tetrameric complex for the retinol binding protein, and within the cavity of the tetramer for thyroxin. The following image shows these active sites,

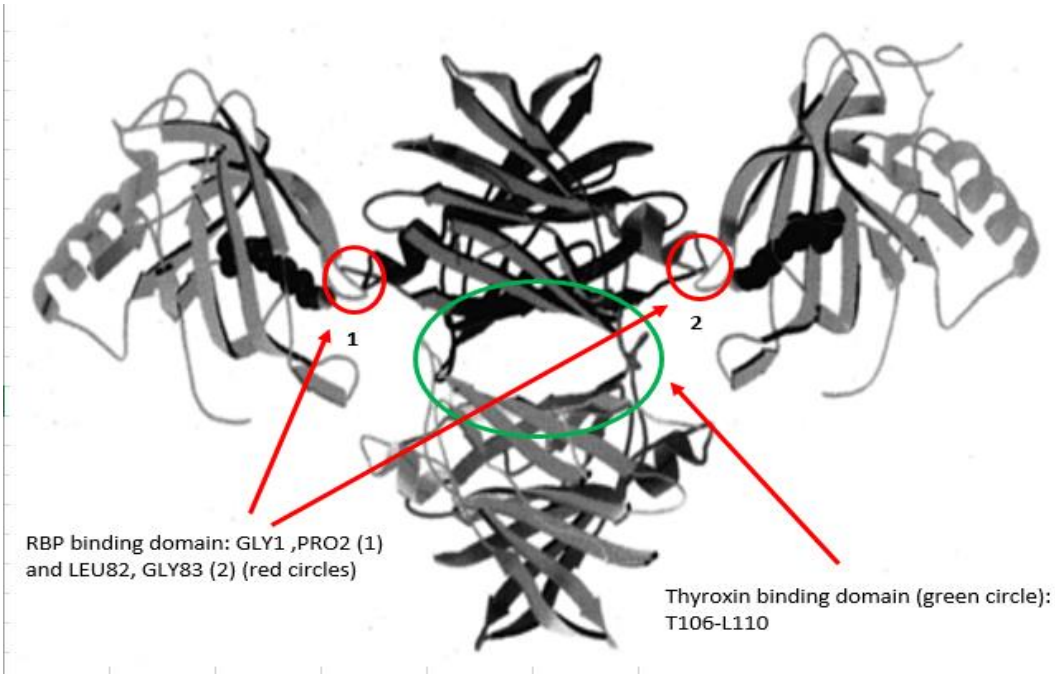


Fig. 8: Retinol binding protein binding sites (red) and the thyroxin active site (green). The retinol binding proteins are on the right and left of the transthyretin complex with retinol bound and shown as a space filling model. The thyroxin binding site is shown empty.

When this protein is defective due to specific residue mutations, a serious condition known as Familial Amyloid Polyneuropathy (FAP) can occur (also known commonly as Corino de Andrade's Disease, named after the Portuguese neurologist who first described it). FAP is an autosomal dominant neurodegenerative disease characterized by pain, parathesia (pins-and-needles sensations), muscular weakness and autonomic dysfunction (Cendron, Trovato, Seno, & et al., 2009). As the scientific name of this disease suggests, FAP is an amyloidogenic disorder, (like Alzheimer's and Parkinson's diseases) wherein amyloid plaque builds up in the brain and results in a general loss of motor function (Cendron et al., 2009). In FAP, amyloid plaques of the mutated protein accumulate in the peripheral nervous system causing a progressive sensory and

motor polyneuropathy (Cendron et al., 2009). Polyneuropathies are damage to multiple nerves in roughly the same area on both sides of the body that lead to weakness, numbness, pins-and-needles sensations (parathesia) and burning pain. The heart and kidneys are affected in the terminal state of this disease (Ando & Jono, 2008). The mutant strain of the transthyretin protein analyzed for this study was a missense mutation resulting in the replacement of Valine by Methionine at the 30th position (V30M) which was not located near any of the active sites (see Figure A10). This is the mutation that is most commonly found in FAP cases (accession number 3DJT) (Cendron et al., 2009). Wild type transthyretin has a half-life of roughly two days in the body (Ando & Jono, 2008). However no information could be gathered for the mutant version of the protein which would have provided useful insight relating the persistent amyloid deposits to the time they reside in the body. OpenContac© was thus performed to analyze the inter-protein chain interactions for this system using Chain A of 3DJT mapped against Chain B of 3DJT. Only these inter-protein interactions were analyzed, not the binding of the ligands of thyroxin or the retinol binding protein.

The so-called "Fine Parsing" results for the transthyretin mapping indicated that the mutant strain exhibited greater interaction from the Lennard-Jones criterion than the wild-type (Table 1). From the total Coulombic interaction and the average distance criteria however, the differences between the two strains were not significant. Note that "UL TOT" and "UCOUL TOT" are the sum of all Lennard-Jones and Coulombic atom pair interactions respectively from the Fine Parsing output.

Table 1: Fine Parsing Results for Transthyretin

	WT	MUT	%
UL TOT:	-5.41E+01	-6.36E+01	15%
UCOUL TOT:	-2.38E+02	-2.46E+02	4%
DISTANCE (AVG):	6.07E-01	5.82E-01	4%

Statistically significant values are hi-lighted: red for potentials and blue for distances. Thus overall, the two strains for this system seem to show little structural variance between them as far as their fine parsing inter-chain interactions were concerned. Visually, this can be seen by comparing the fine parsing images (chains have been hi-lighted and labeled),

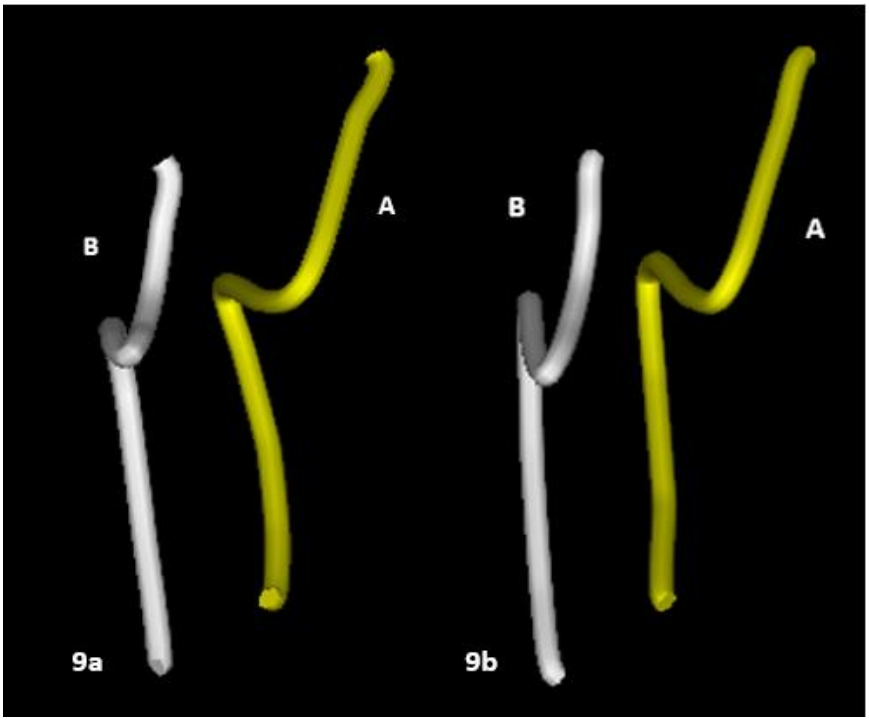


Fig. 9: transthyretin wild-type (9a) and mutant (9b) fine parsing

Wild-type residue range: 33-122 (A chain), 41-122 (B chain)

Mutant residue range: 39-122 (A chain), 39-122 (B chain)

By considering the shift in Lennard-Jones interactions when going from the wild-type strain to the mutant strain, one can see that the mutation permits a novel situation wherein the non-charged, non-polar residues are now interacting stronger in the mutant strain. Non-charged and non-polar residues are usually associated with the interior region, or endo-domains, of a protein and are hydrophobic in nature so the conclusion from the overall analysis of the protein is that its interior has stronger hydrophobic interactions in the mutated form. The schematic representation was used for fine parsing visualization, as it permits visualization of the most basic alpha carbon backbone which is very often lost with ribbon representation. This paradigm will be used throughout this study.

From a structural perspective, the coarse parsing also seemed to show overall retention of structure (Fig. 10).

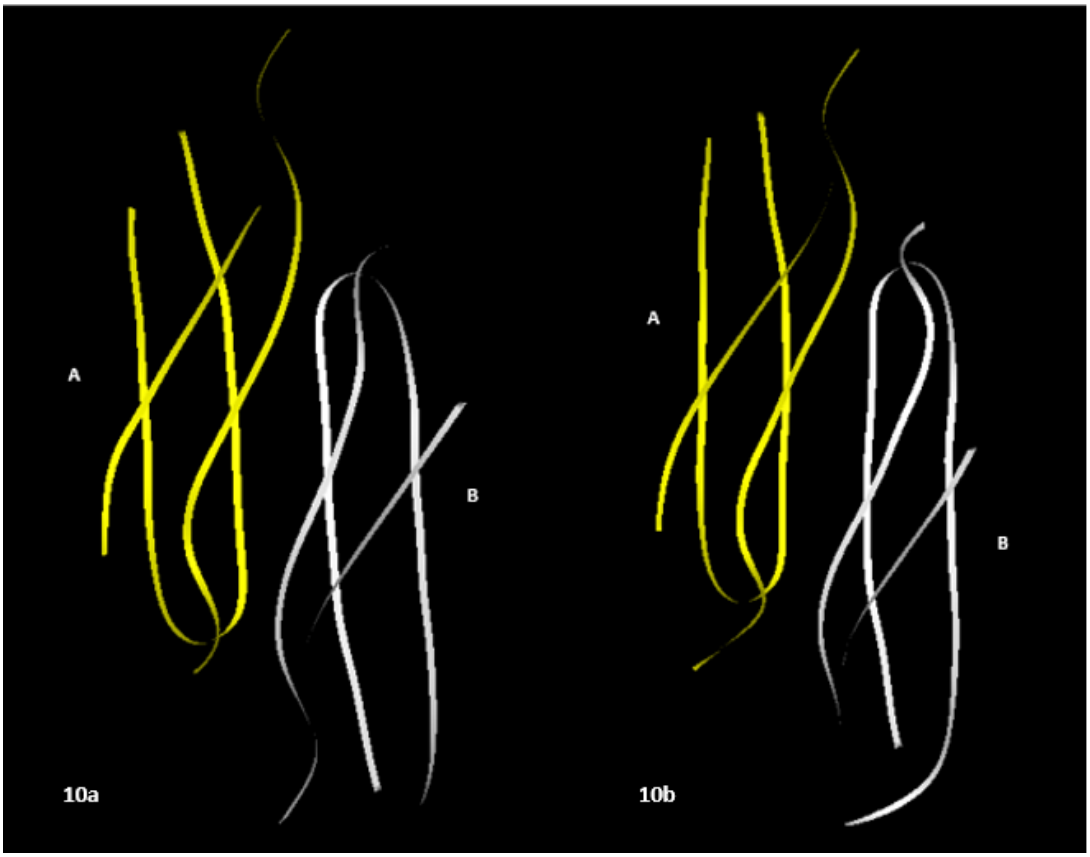


Fig. 10: Transthyretin wild-type (10a) and mutant (10b)

Wild-type residue range: 12-124 (A chain), 11-124 (B chain)

Mutant residue range: 11-124 (A chain), 11-123 (B chain)

In order to further quantify changes in protein-protein interaction (PPI) behavior, common and un-common atom-atom interactions between the two strains were identified from OpenContact© Fine Parsing spreadsheets. The common interactions were atom-atom pairs found in both wild-type and mutant strains whereas the un-common interactions represent those atom-atom pairs that were unique to each strain. For this analysis, virtually every residue present in the whole fine parsing output had a constituent atom participate in common interactions (i.e. almost all residues are common residues). The results for the common atom-atom interactions are shown in the Table 2.

Table 2: Common Results for Transthyretin

	WT	MUT	%
UL TOT:	-3.68E+01	-3.59E+01	2%
UCOUL TOT:	-2.12E+02	-2.24E+02	5%
DISTANCE (AVG):	5.94E-01	6.03E-01	1%
COMMON ATOM-ATOM PAIRS:	543		

For these interaction pairs, one can see that there was little deviation between the wild-type and mutant strains. Intuitively, this makes sense as the atom-atom interaction pairs considered are identical in their participants between the wild-type and mutant with very little overall structural changes. The analysis of the un-common interactions thus considered the remaining atom-atom pairs that were unique to each individual strain. The results for this analysis are shown in Table 3.

Table 3: Un-common Results for Transthyretin

	WT	MUT	%
UL TOT:	-1.72E+01	-2.77E+01	38%
UCOUL TOT:	-2.64E+01	-2.13E+01	19%
DISTANCE (AVG):	6.35E-01	5.46E-01	14%
UN-COMMON ATOM-ATOM PAIRS:	258	307	

By identifying only the un-common interaction partners, a clear shift was seen going from Coulombic to Lennard-Jones when going from the wild-type to the mutant. Additionally, the average inter-atomic distance decreased slightly as visually evident by overlaying Fig. 9a over Fig. 9b (not shown). It is interesting to note that despite these findings, the biology of the disease

state still permits transport of the two ligands (thyroxin and the retinol binding protein) (Cendron et al., 2009). This is consistent with the lack of overall structural changes. How the subtle change or shift in PPIs affects plaque formation is not known, but these PPI changes have been clearly detected here.

5.1 SYSTEMS ANALYZED: SOD1

The next system that was analyzed was the enzyme SOD1. The biologically active form of this enzyme is a homo-dimer with each monomer having a molecular weight of 15.94 kDa for a total molecular weight of 31.88 kDa for the whole protein ("SOD1 (human)," 2014). Below is an image of the wild type and mutant strains that were used for this study (accession numbers 4FF9 and 3GZQ respectively),

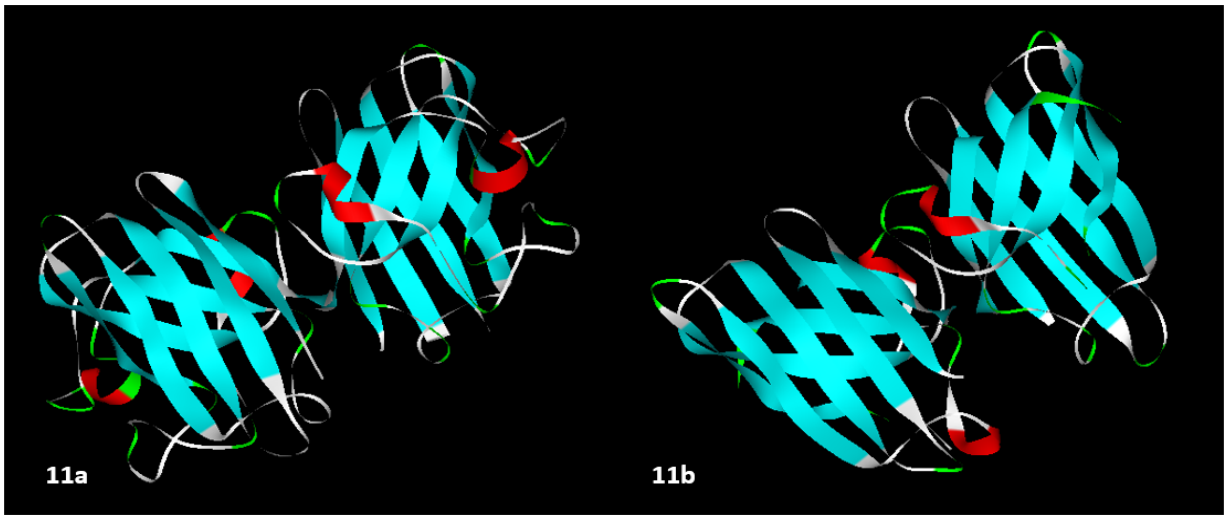


Fig. 11: SOD1 dimer enzyme wild-type (11a) and mutant (11b). Accession numbers 4FF9 and 3GZQ respectively

Superoxide dismutase 1, or SOD1, is an enzyme that is responsible for destroying superoxide radicals in the human body (Auclaire, Brodkin, D'Aquino, Petsko, Ringe & Agar, 2013). It belongs to a general family of enzymes known as superoxide dismutases that catalyze the dismutation of superoxide (O_2^-, a free radical) into diatomic oxygen and hydrogen peroxide thus making the enzyme an extremely important anti-oxidant (Auclaire et al., 2013). The presence of superoxide in biological systems, specifically in humans, comes from the by-products of mitochondrial respiration as well as from the immune system which specifically manufactures superoxide for use in oxygen-dependent killing mechanisms of foreign invading pathogens (Auclaire et al., 2013). Being a free radical, superoxide is extremely toxic due to its high reactivity and can readily react with (and therefore alter) DNA. The neutralization of superoxide is accomplished by the presence of metal ions in an accessible region of the protein (its active site, though more on this later) that consequently makes SOD1 and the other SOD proteins metalloproteins (Auclaire et al., 2013). The three families are distinguished from one another according to the specific metal/metals present in the active site, which in turn dictates the location of activity (i.e. intracellular or extracellular) or the specific types of organisms serviced by the dismutase (Galaleldeen & Strange, 2009). The subject of study for this project was the human superoxide dismutase known as SOD1 which is found in the cytoplasm of cells and contains copper and zinc as its active site metal ions, both of which are in the +2 oxidation state (Ramirez, Mejiba Gomez, & Mason, 2005). The other two classes of SOD, SOD2 (which uses iron or manganese) and SOD3 (which uses nickel) are also found in humans in the mitochondria and extracellular fluids respectively (Selverstone Valentine, Doucette, & Zittin Potter, 2005). Prokaryotes will either use SOD2 or SOD3 for their superoxide neutralization protocols (Selverstone Valentine et al., 2005).

Each individual monomer in SOD1 is built upon a β barrel motif with one copper and one zinc binding site. Two large loops called the electrostatic and zinc loops encase the metal-binding region which serves as the active site (although the heart of enzymatic activity revolves around the copper ion). Thus the whole biologically active SOD1 enzyme has a total of two active sites, one on each monomer. Below is the entire enzyme with the metal ions in the active site hi-lighted as reported by Selverstone Valentine et al.,

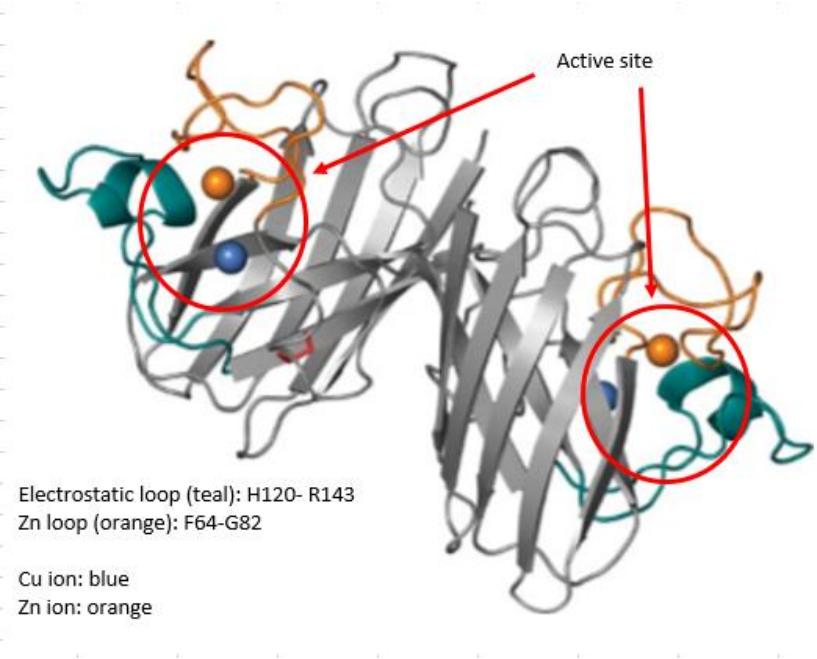


Fig. 12: SOD1 enzyme (wild-type) with loop regions and active site

The overall enzymatic activity of SOD1 is described in the following chemical reaction equation as given by Selverstone Valentine et al.,

(11) $2O_2^- + 2H^+ \rightarrow O_2 + H_2O_2$

where two hydronium ions (technically shown here as free protons) from the aqueous *in vivo* environment provide the necessary hydrogens to evolve hydrogen peroxide. Considering the detailed chronological mechanism by which the cupric ion participates involves a two-step reaction process. The first step is shown next,

(12) $O_2^- + Cu^{2+}ZnSOD \rightarrow O_2 + Cu^{+}ZnSOD$

Here, one molecule of superoxide first reduces the cupric ion to the cuprous oxidation state (+1) to form diatomic oxygen. The second reaction that yields the final hydrogen peroxide product is shown next,

(13) $O_2^- + 2H^+ + Cu^{+}ZnSOD \rightarrow H_2O_2 + Cu^{2+}ZnSOD$

Thus the copper ion is returned to the cupric oxidation state thereby regenerating the enzyme for further activity.

Should a mutation occur, a disease known as amyotrophic lateral sclerosis or ALS (known as Lou Gehrig's disease in the United States) can develop. ALS is a neurodegenerative disease characterized by rapidly progressing weakness due to muscle atrophy, muscle spasticity, difficulty speaking, swallowing and breathing (Galaleldeen & Strange, 2009). Although the disease can be inherited, this mode of acquisition only applies to about 5%-10% of cases (thus termed familial ALS) with the majority of incidents occurring spontaneously (Galaleldeen & Strange, 2009). These symptoms are brought about by the gradual degeneration of the upper and lower motor neurons and although all voluntary movement is ultimately lost, bladder and bowel

functions as well as the muscles responsible for eye movement are typically spared until the final stages of the disease (Galaleldeen & Strange, 2009). Interestingly, cognitive function is generally spared for most ALS patients, although some 5% of sufferers may develop frontotemporal dementia (Galaleldeen & Strange). Survival time with this disease is around 39 months from the onset of symptoms and death is ultimately caused by respiratory failure. A well-known case of ALS is that of cosmologist Stephen Hawking who first started displaying symptoms during his years at Cambridge University as a doctoral student. He has since lived with this disease for more than 50 years, however his case is exceptional. The exact mechanism by which patients experience a gradual loss of neuronal function is as of yet still unknown, but aggregates of large fibrils of proteinaceous inclusions rich in mutant SOD1 are characteristic of tissue samples from ALS patients (Selverstone Valentine et al., 2005). As has been proposed for ALS and other neurodegenerative diseases, the fibrils themselves may not be the toxic species as they are formed relatively late in the disease, but instead may be a final accumulation of the insoluble protein form as the body becomes incapable of handling such a large buildup. This scenario might therefore lead one to believe that the half-life of the mutant strain is longer than that of the wild-type since fibrils of mutated protein are found in the aggregates. Interestingly enough, this is not the case. According to Selverstone Valentine et al., wild-type human SOD1 has a half-life of 30 hours. This is dramatically different from the mutant's half-life (specifically the mutant that was studied) which is 7.5 hours. The mutant strain utilized in this study was the A4V mutation, accession number 3GZQ.

The overall fine parsing analysis for SOD1 yielded results that were by and far too close between the wild-type and mutant strains to be considered favoring either. Below is the table summarizing these results,

Table 4: Fine Parsing Results for SOD1

	WT	MUT	%
UL TOT:	-3.99E+01	-4.07E+01	2%
UCOUL TOT:	-2.34E+02	-2.11E+02	10%
DISTANCE (AVG):	6.55E-01	6.43E-01	2%

Thus the data implies a striking similarity in binding interactions between both wild-type and mutant amongst all of the atom-atom interactions responsible for keeping the two sub-units together. The pdb molecular representation for this fine parsing data is shown below,

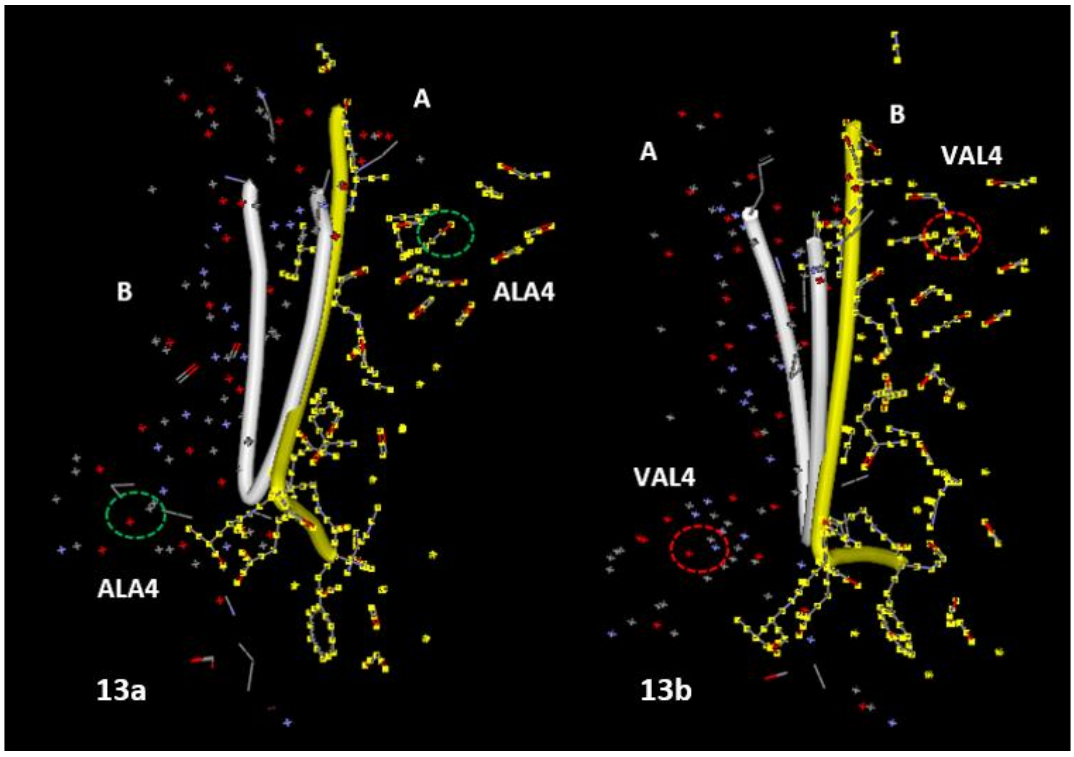


Fig. 13: SOD1 wild-type (13a) and mutant (13b) fine parsing

Wild-type residue range: 3-153 (A chain), 1-153 (B chain)

Mutant residue range: 1-153 (A chain), 3-153 (B chain)

Positions that underwent mutations are circled in Figures 15 and 16 as this will be the paradigm for all the images that are to follow, should there be a mutation present in the fine or coarse parsing output images. One item of note to point out is that the original pdb input data file for the wild-type and mutant had the A and B chain nomenclature reversed. This was of no consequence however for OpenContact© as the results yielded are not dependent on chain name.

An interesting aspect of this case was that the residue position (the 4th position) that was mutated also participated in inter-chain binding as seen in Figures 15 and 16 (also recall Figure A11 which shows this residue position near the inter-chain region). The specific differences between the interactions for the mutated position alone is shown in the table below,

Table 5: SOD1 Mutant Position Comparison (4)

	WT	MUT	%
UL TOT:	-7.53E-02	-5.21E-02	30%
UCOUL TOT:	-6.12E+00	-3.56E+00	42%
DISTANCE (AVG):	7.62E-01	7.79E-01	2.00%

Although the mutated residue position was present in the fine parsing studies for both strains, it is crucial to point out that not all of the same atom-atom binding pairs participated in binding. As it turned out, the wild-type had fourteen of its constituent binding pairs participate whereas the mutant only had ten. Nonetheless, the wild-type's binding pairs exhibited considerably stronger interactions as seen in Table 5. Interestingly enough, despite superior binding from the wild-type, the average distance for both strains were almost identical.

With respect to the structural changes, the mutant strain exhibited some considerable change from the wild-type and even showed the development of secondary structure in the form of an anti-parallel beta sheet. These results are shown below,

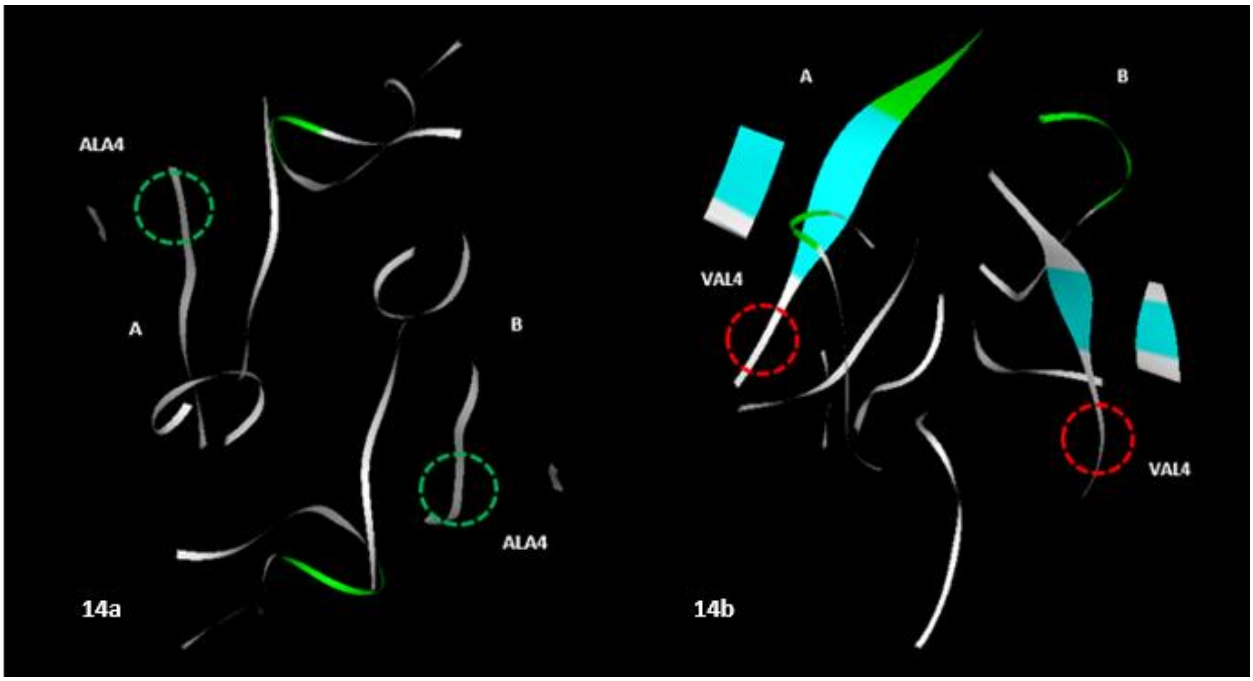


Fig. 14: SOD1 wild type (14a) and mutant (14b) coarse parsing

Wild-type: 1-153 (A chain), 1-153 (B chain)

Mutant: 1-153 (A chain), 2-153 (B chain)

As can be seen, there is a significant change in structure in the mutant, particularly a noticeable introduction of beta sheet formation in the mutant as well as a re-orientation of sorts of the mirror image symmetry of the two mutant chains. Where the wild-type has inverted mirror image symmetry (with each mutant position on opposite ends), the mutant is a very good mirror image of the two constituent chains. More striking is how, as a result of this, the mutated position is now very much in symmetry in the mutant strain. As in the fine parsing results, the

mutated residue shows up in both strains and is noticeably closer to the interior of the binding region in the mutant compared to the wild-type.

When considering the atom-atom interactions common to both the wild-type and mutant strains, a similar trend of virtually equal interactions across all three criteria is observed. The following table summarizes these results,

Table 6: Common Results for SOD1

	WT	MUT	%
UL TOT:	-3.29E+01	-3.26E+01	1%
UCOUL TOT:	-2.00E+02	-1.99E+02	1%
DISTANCE (AVG):	6.44E-01	6.42E-01	1%
COMMON ATOM-ATOM PAIRS:	606		

As would be expected, there was practically no change of interaction type when going from the wild-type to the mutant strain. As was observed for transthyretin, these common interactions pertained to virtually every residue from the fine parsing output.

Upon examination of the un-common interactions however, one can see that there begin to be signs of bias in favor of the wild-type strain or the mutant strain depending on the interaction criterion considered. These results are summarized below,

Table 7: Un-common Results for SOD1

	WT	MUT	%
UL TOT:	-6.91E+00	-8.10E+00	15%
UCOUL TOT:	-3.42E+01	-1.19E+01	65%
DISTANCE (AVG):	7.00E-01	6.47E-01	8%
UN-COMMON ATOM-ATOM PAIRS	148	108	

The results indicate that the mutant strain contains greater Lennard-Jones interactions, again strongly associated with non-polar residues, and that the wild-type presents an even greater proportion of electrostatic activity associated with polar residues.

5.2 SYSTEMS ANALYZED: P53

The next system analyzed was p53, a protein highly involved in DNA repair and programmed cell death (apoptosis) that is strongly linked with incidences of cancer. The biologically active form of P53 protein is that of a tetramer complex composed of four identical subunit chains. Each chain has a mass of 43.7 kDa for a total mass of 174.8 kDa ("p53 (human)," 2014). Below is an image of wild-type p53 as it would be found executing its DNA reparative function as reported by Malecka, Ho, & Marmorstein, 2009,

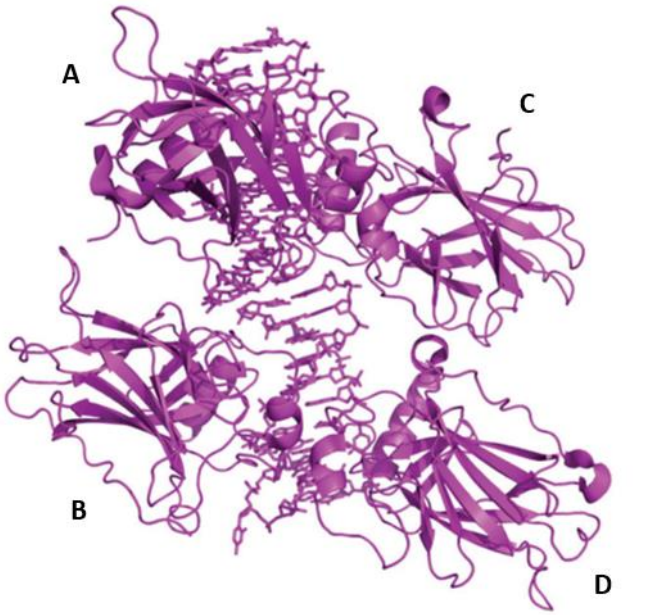


Fig. 15: p53 tetramer complexed to DNA

Although the above image is of the whole protein, it must be noted that crystallizing the entire protein is difficult due to the inherent flexibility of p53 that is part of its biological activity and function (more on this later) (Mora & Carbajo, 2008). Thus the structure pdb file available for study was only of the tetramer's binding domains. The wild type strain employed for this study was of the accession number 2J0Z (Mora & Carbajo, 2008).

p53 is an important protein involved in repairing DNA damage and is thus extremely important in cancer cell systems (Goodsell, 2014). Over 70% of all cancers exhibit some form of p53 mutations regardless of whether or not the mutations are acquired post-nattily or inherited (Gao et al., 2013). In general, mutations that lead to cancer abide by either one of two mechanisms: a mutation that results in uncontrolled growth/multiplication of cells or mutations that block the normal defenses responsible for protecting against unnatural growth (Goodsell, 2014). Consequently, cancerous cells will exhibit mutations that affect either one (or possibly

even both) of these regulatory processes. In the case of p53 the latter of the two mechanisms (unnatural growth protection) is regulated. P53 is normally found in low levels, however, should DNA damage be detected, the levels of this protein would rise and initiate the production of proteins that halt cell division until the damage is repaired (Mora & Carbajo, 2008). If the damage is too severe, however, p53 initiates the process of programmed cell death, or apoptosis, which ultimately directs the cell to commit suicide thus permanently removing it from the in-vivo system (Mora & Carbajo, 2008). When p53 executes its function, it binds to regulatory sites in the genome that in turn trigger the production of the aforementioned proteins such as WAF1 and G1-S/CDK (Mora & Carbajo, 2008). In the case of WAF1 and G1-S/CDK, these two proteins complex together which results in the affected cell being unable to proceed with the next stage of cell division (Mora & Carbajo, 2008). As the name G1-S/CDK implies, cell growth is usually halted at the G1/S phase by the complex that this protein participates in (Mora & Carbajo, 2008). Due to the enormous benefit provided by p53, it is commonly given the nick-name of tumor suppressor protein and guardian of the genome. In all, the role p53 plays in neutralizing rogue cells is tantamount to the role of helper T cells in the immune system: both identify harmful pathogens (the immune system "foreign" and p53 "domestic") and both invoke entities to do the actual neutralization of the target species.

As stated earlier, p53 binds to target regions/section of DNA that direct the transcription and ultimate translation of proteins such as WAF1 and G1-S/CDK. Physically, this is accomplished by the p53 tetramer opening up from its normally closed binding domain and binding to the appropriate location on the gene (Mora & Carbajo, 2008). It is important to point out however, that the structure to follow in Figure 20 does not bind to DNA. Figure 16 shows the structure responsible for p53 chains binding to and amongst themselves for both the wild-

type and mutant strains. Malecka et al. reported the precise DNA binding residues (NOT the tetramer binding region) as being the 99th residue to the 289th. The residues in Figure 16 (the binding region studied in this work) span from 326th to the 356th residues. Thus below is an image of the wild-type and mutant binding domains of the protein for inter-chain, intra-protein stability which was also incidentally the structure used for this study (the chains are also labeled). The accession number for the wild-type is 2J0Z and for the mutant it is 2J10.

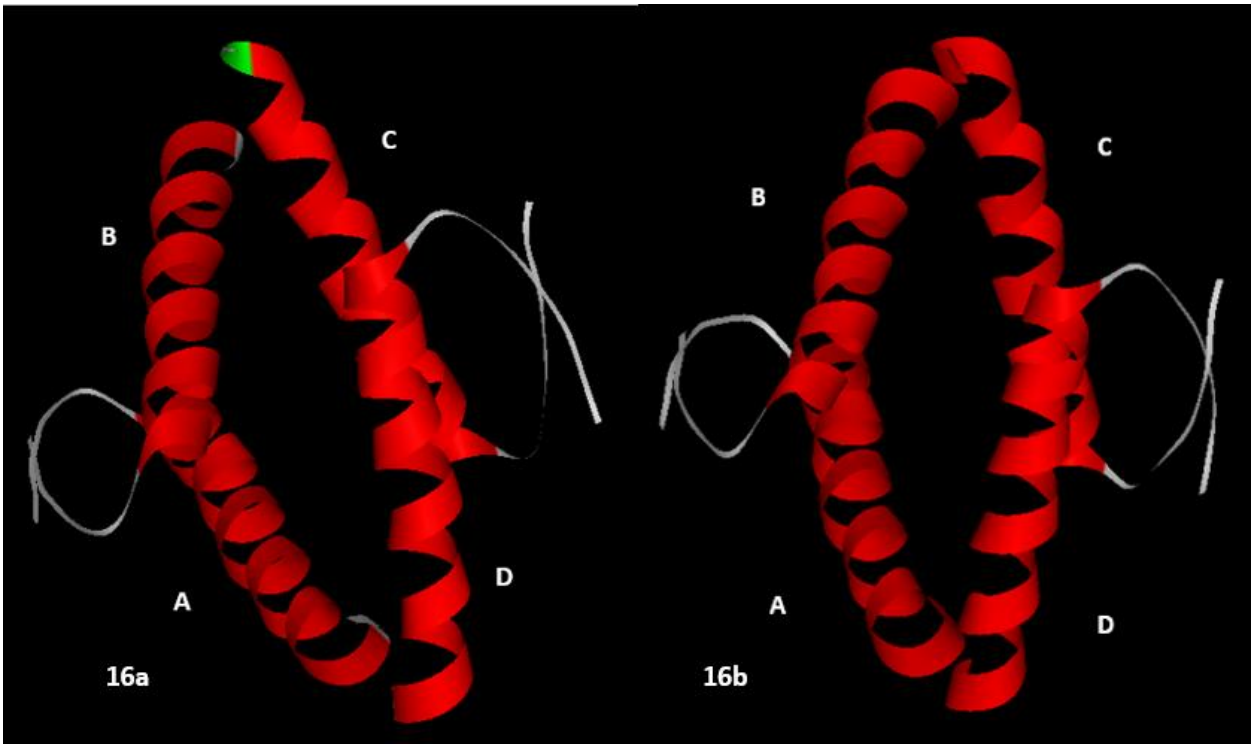


Fig. 16: p53 binding domain, wild-type (16a) and mutant (16b). Accession numbers: 2J0Z and 2J10 respectively.

As can be seen, the tetramerization binding for this protein is a system of four alpha helices. Each helix corresponds to and is attached to one of the monomers that make up the rest of each monomer structure. As briefly mentioned earlier, the protein is flexible and it is in the binding

domain that the flexibility is extremely important as this is the region wherein p53 binds to DNA. This is unfortunately a problem for crystallographers as flexible proteins do not tend to form orderly crystals, and if they do crystallize, the results are often of a low resolution (Malecka et al., 2009). Structural determination has thus been mostly elucidated by piece-wise studies on the protein, one of which is the binding domain featured in this study.

Mutations in p53 have a negative impact in the protein's ability to carry out its function, particularly, its immediate ability to effectively bind to DNA. The outcome of this impaired reparatory ability is ultimately cancer. The human body is constantly being bombarded with materials such as chemicals, ionizing radiation, that all have the potential to interact with DNA and damage it (recall superoxide from mitochondrial respiration specially handled by SOD1). A defective p53 will therefore not be able to invoke the proteins such as WAF1 and G1-S/CDK that halt an affected cell thus allowing the growth of a tumor (Mora & Carbajo, 2008). In executing its function it is noteworthy to mention that the p53 DNA binding domains interact primarily with the major grooves of the DNA molecule (see Figure 19). The mutant strain that was ultimately used sported two mutations at the 329th (T329F) and 331st (Q331K) positions not in the DNA binding domain and was of the accession number 2J10 (Mora & Carbajo, 2008).

One of the unique scenarios presented by this particular case was its multiplicity of chains (more than two) that required analysis. This presented an opportunity to study the binding interactions present in the different combinations of chains to see what, if any trends could be found. Thus the first module for p53 analysis focuses on the A-X chain family or combinations (i.e. A-B, A-C and A-D). The other combinations will be considered later in their own separate module reviews. For the overall Fine Parsing PPI analysis, the results are shown below,

Table 8: Fine Parsing Results for p53, A-B

	WT	MUT	%
UL TOT:	-5.35E+01	-4.25E+01	21%
UCOUL TOT:	-3.04E+02	-2.60E+02	14%
DISTANCE (AVG):	5.89E-01	6.08E-01	3%

Table 9: Fine Parsing Results for p53, A-C

	WT	MUT	%
UL TOT:	-1.56E+00	-1.10E+00	30%
UCOUL TOT:	-2.66E+01	-3.59E+01	26%
DISTANCE (AVG):	8.04E-01	8.03E-01	0.1%

Table 10: Fine Parsing Results for p53, A-D

	WT	MUT	%
UL TOT:	-5.31E+00	-5.40E+00	2%
UCOUL TOT:	-6.87E+01	-7.81E+01	12%
DISTANCE (AVG):	6.97E-01	7.11E-01	2%

As can be seen, the overall Fine Parsing analysis yielded results that were relatively intuitive given the structure of the binding domain. The A-B interactions (Table 8) yielded the greatest total interaction quantities and smallest average separation distances among all the other interaction pairs (A-C, A-D). The difference was one order of magnitude greater for the A-B interactions compared against both A-C and A-D potentials. Average distances between wild-type and mutant strains, however, remained in the same order of magnitude for all three A-X

interactions. One can see by visual inspection of Fig. 20 that in the overall structure, the greatest inter-chain overlap indeed occurs between the A-B chains and, thus, the data corroborates that these two chains certainly have the strongest interaction potentials. "Second place" in total interaction potential went to the A-D chain interaction system with the next largest potentials and smallest separation distances. Needless to say, A-C by far had the weakest interactions. When tracking the shifts in potentials across the A-X chain family, we observe that there is an overall decrease in in both Lennard-Jones and Coulombic potentials when going from the wild-type to the mutant for the A-B chain interaction system. For the A-C pair, we see that there was a shift from Lennard-Jones to Coulombic likewise when going from the wild-type to the mutant and for the A-D chain interaction system there was an increase in Coulombic interactions.

Considering the associated Fine Parsing structure for these three interaction scenarios, one can see that the mutated positions show up most predominantly in the A-B chain interactions (Fig. 17). We can thus see that the mutated residues play an important role in binding for the mutant strain and may account for the behavior noted the preceding discussion. The fine parsing images for these results follow,

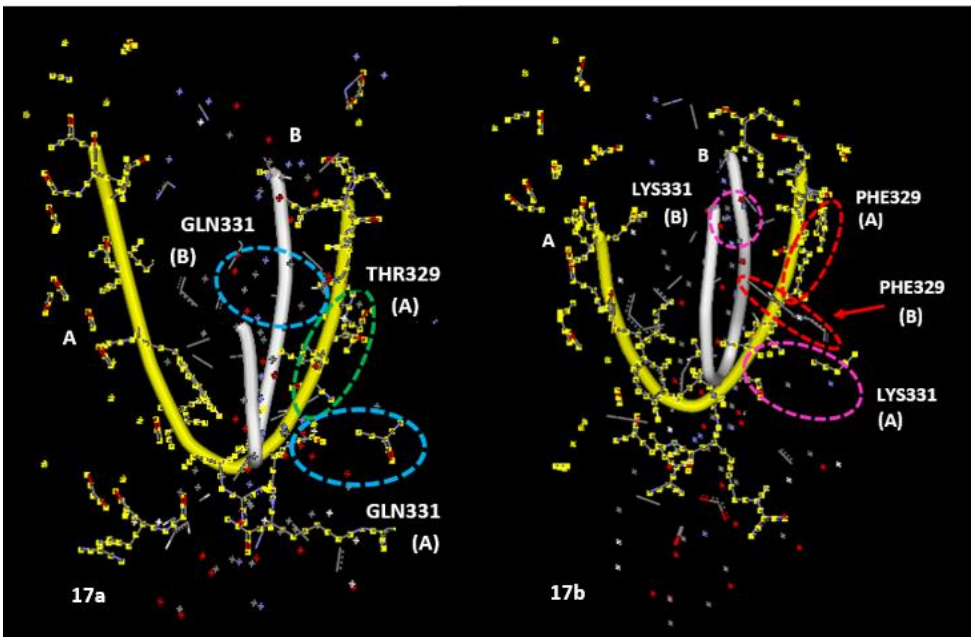


Fig. 17: p53 wild-type (17a) and mutant (17b) Fine Parsing, A-B

Wild-type: 326-352 (A chain), 326-352 (B chain)

Mutant: 326-356 (A chain), 326-352 (B chain)

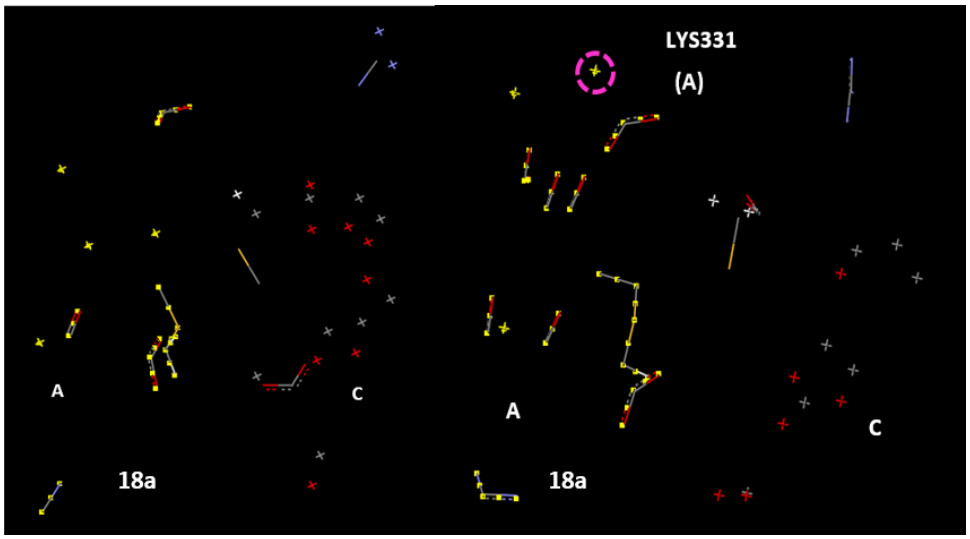


Fig. 18: p53 wild-type (18a) and mutant (18b) Fine Parsing, A-C

Wild-type: 336-344 (A chain), 336-344 (C chain)

Mutant: 335-343 (A chain), 335-343 (C chain)

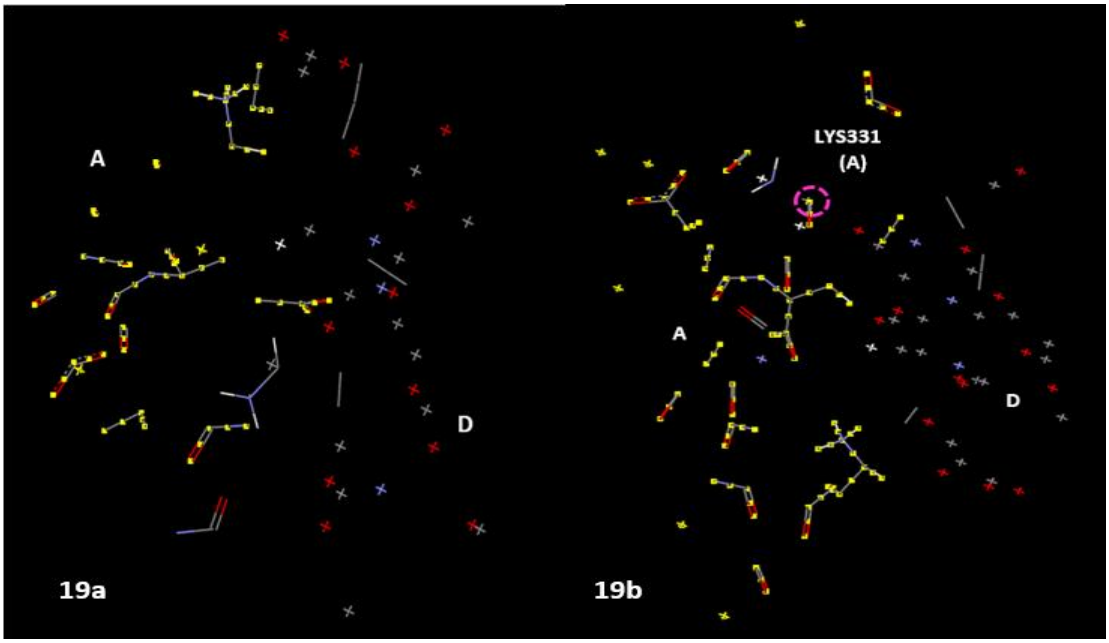


Fig. 19: p53 wild-type (19a) and mutant (19b) Fine Parsing, A-D

Wild-type: 343-354 (A chain), 343-354 (D chain)

Mutant: 339-355 (A chain), 342-354 (D chain)

From analyzing the above images, one can see that the A-B interaction system clearly had more atoms participate in binding (so much that some degree of structure was even identified by DS Viewer Pro©), thus accounting for the stronger overall interactions observed for the A-B system. This stands in stark contrast to the fine parsing images for the A-C and A-D interactions that clearly have fewer atoms participating in their binding and do not have any recognizable associated structure. Additionally, one can see that atoms from both mutated residue positions (the 329th and 331st residues) contributed to A-B interactions for both the wild-type and the mutant, but only one select atom from the mutation at the 331st residue played any role in the A-C and A-D binding and only for the mutant strain.

Upon examination of the coarse parsing results, one can see that structurally, there is little difference between the wild-type and the mutant. Although, as will be seen, there are some slight differences with actual atomic orientations and angles, these differences manifest themselves closer to the binding intra-protein, inter-chain domain of the tetramer and do not appear to be overwhelmingly impactful on the overall morphology of the protein. Below are images taken from these results,

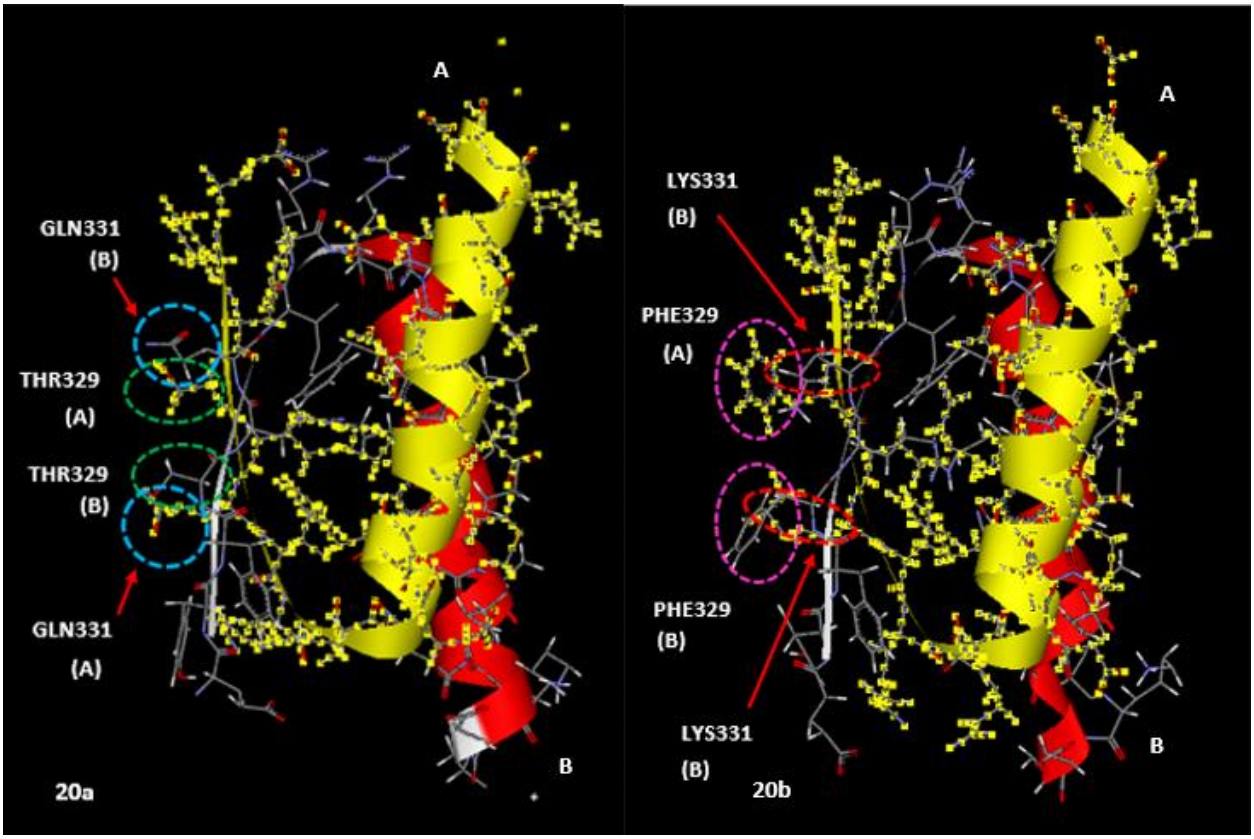


Fig. 20: p53 wild-type (20a) and mutant (20b) coarse parsing, A-B

Wild-type: 326-352 (A chain), 326-352 (B chain)

Mutant: 326-356 (A chain), 326-352 (B chain)

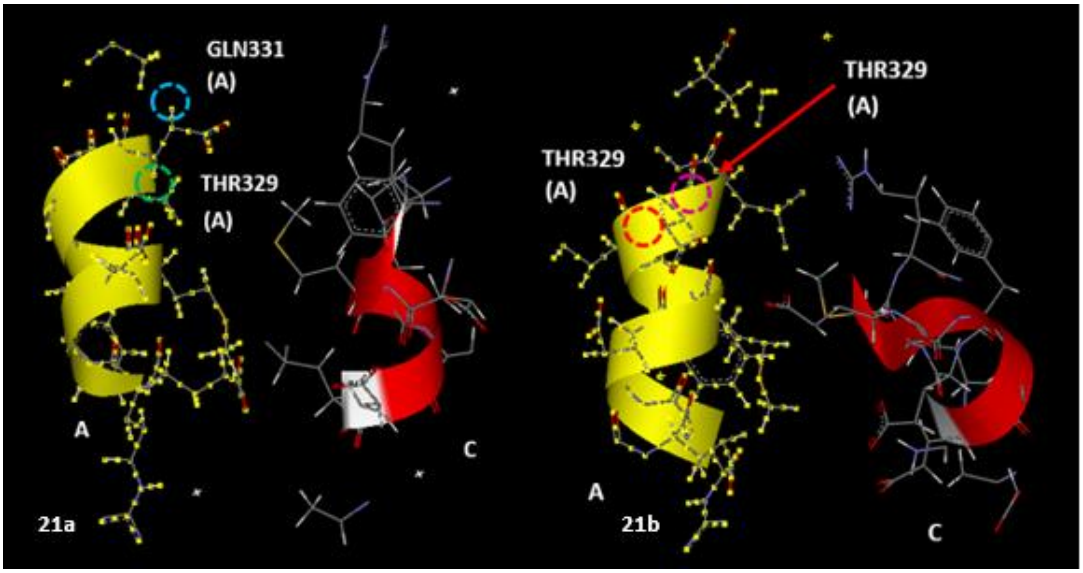


Fig. 21: P53 wild-type (21a) and mutant (21b) coarse parsing, A-C

Wild-type: 336-348 (A chain), 334-351 (C chain)

Mutant: 330-351 (A chain), 335-351 (C chain)

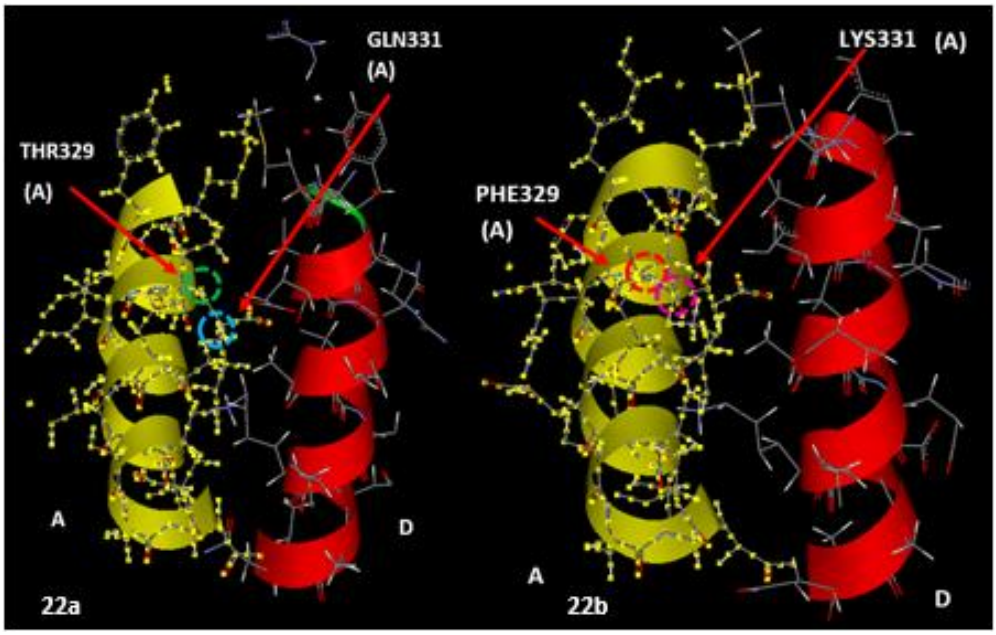


Fig. 22: P53 wild-type (22a) and mutant (22b) coarse parsing, A-D

Wild-type: 337-355 (A chain), 336-355 (D chain)

Mutant: 336-356 (A chain), 337-355 (D chain)

When analyzing the common residue component of this system (common atom-atom pair interactions for the A chain and X chain interactions), there are slight shifts in the total common interaction potential values (Tables 14-16) with a slight weakening in A-B mutant interactions over the wild-type strain. Chain interactions A-C and A-D experience a shift in Coulombic and Lennard-Jones interactions respectively when going from the wild-type to the mutant.

Table 14: Common Results for p53 (A-B)

	WT	MUT	%
UL TOT:	-2.12E+01	-1.83E+01	14%
UCOUL TOT:	-2.14E+02	-1.78E+02	17%
DISTANCE (AVG):	6.07E-01	6.37E-01	5%
COMMON ATOM-ATOM PAIRS:	428		

Table 15: Common Results for p53 (A-C)

	WT	MUT	%
UL TOT:	-6.80E-01	-7.02E-01	3%
UCOUL TOT:	-2.24E+01	-2.55E+01	12%
DISTANCE (AVG):	8.27E-01	7.78E-01	6%
COMMON ATOM-ATOM PAIRS:	42		

Table 16: Common Results for p53 (A-D)

	WT	MUT	%
UL TOT:	-1.67E+00	-2.13E+00	22%
UCOUL TOT:	-5.43E+01	-4.90E+01	10%
DISTANCE (AVG):	7.19E-01	7.22E-01	0.4%
COMMON ATOM-ATOM PAIRS:	105		

However, for the uncommon interaction partners, the result is a dramatic shift from Lennard-Jones to Coulombic interactions which on the whole is consistent with residue side chain behavior given that the mutant has a Phenylalanine (non-polar) replacing a Threonine (polar) and a Lysine (charged) replacing a Glutamine (polar). Thus in the end, the mutations provide charged species that were lacking in the wild-type.

Table 17: Un-common Results for p53 (A-B)

	WT	MUT	%
UL TOT:	-3.22E+01	-2.42E+01	25%
UCOUL TOT:	-9.01E+01	-8.15E+01	10%
DISTANCE (AVG):	5.74E-01	5.76E-01	0.4%
UN-COMMON ATOM-ATOM PAIRS:	490	396	

Table 18: Un-common Results for p53 (A-C)

	WT	MUT	%
UL TOT:	-8.76E-01	-4.03E-01	54%
UCOUL TOT:	-4.20E+00	-1.04E+01	60%
DISTANCE (AVG):	7.42E-01	8.41E-01	12%
UN-COMMON ATOM-ATOM PAIRS:	16	27	

Table 19: Un-common Results for p53 (A-D)

	WT	MUT	%
UL TOT:	-3.64E+00	-3.27E+00	10%
UCOUL TOT:	-1.44E+01	-2.91E+01	51%
DISTANCE (AVG):	6.54E-01	7.00E-01	7%
UN-COMMON ATOM-ATOM PAIRS:	54	93	

Given the participation of the mutated residues, it is interesting to note how the interaction quantities compare from the wild-type to the mutant. Since the mutated position was present in both the wild-type and mutant strains in the fine parsing of the A-B interaction system, the following analysis was carried out. This was not the case with A-C and A-D however which had a mutant residue position only show up in the mutant strain. The following three tables present these findings,

Table 11: p53 Mutant Position Comparison (329 & 331)

	WT	MUT	%
UL TOT:	-2.12E+01	-5.83E+00	73%
UCOUL TOT:	-1.44E+02	-5.05E+01	65%
DISTANCE (AVG):	5.77E-01	6.22E-01	7%

As can be seen, interactions are greater in the wild-type for both Lennard-Jones and Coulombic interactions. As observed with the overall analysis previously discussed, the Coulombic interactions are greater within the wild-type and this observation serves as a testament to the nature of the forces that dominate in the endo-domain binding region of p53.

The above analysis results were for the combined effects of both positions that exhibited mutations. Considering now just the 329th position, one can see yet again the bias towards the wild-type strain in having stronger interactions.

Table 12: p53 Mutant Position Comparison (329)

	WT	MUT	%
UL TOT:	-1.80E+01	-2.37E+00	87%
UCOUL TOT:	-1.06E+02	-1.47E+01	86%
DISTANCE (AVG):	5.47E-01	5.73E-1	5%

Interestingly enough, the average inter-atomic distances between wild-type and mutant practically did not change. Even though the wild-type strain clearly had better interactions, once again, we see the dominance of electrostatic forces as being considerably greater than Lennard-Jones forces (electrostatic Coulombic interaction potentials were greater by an order of magnitude).

Looking at the 331st position now, one can see an interesting turn of events for the ranking of inter-atomic binding with the scenario now showing no preference for either strain,

Table 13: p53 Mutant Position Comparison (331)

	WT	MUT	%
UL TOT:	-3.24E+00	-3.46E+00	6%
UCOUL TOT:	-3.85E+01	-3.58E+01	7%
DISTANCE (AVG):	6.71E-01	6.45E-01	4%

Although there are practically no changes when going from one strain to another, one can definitely see that the coulombic electrostatic forces predominate in both strains nonetheless. This is interesting to note due to the fact that glutamine (present in the wild-type) is typically treated and behaves as an uncharged amino acid in contrast to lysine (present in the mutant) which is treated as a charged amino acid due to its side group. Nevertheless, the two amino acids behave very similarly to each other as evidenced by the magnitude of their Lennard-Jones and Coulombic interaction values.

Now that the A-X family of interactions has been described and discussed, we proceed to the next module for consideration of the chain combination: B-X. Since the original combinations for a quaternary chain system apply to only two members for the B-X family (B-C and B-D), there are consequently only two combinations to look at now as opposed to the three that there were for the A-X family (A-B, A-C and A-D). As before, the analysis for this family will begin with the overall calculation results.

Table 20: Fine Parsing Results for p53 (B-C)

	WT	MUT	%
UL TOT:	-3.74E+00	-7.15E+00	48%
UCOUL TOT:	-5.46E+01	-8.33E+01	34%
DISTANCE (AVG):	7.24E-01	6.91E-01	5%

Table 21: Fine Parsing Results for p53 (B-D)

	WT	MUT	%
UL TOT:	-2.48E+00	-1.25E+00	50%
UCOUL TOT:	-2.70E+01	-1.18E+01	56%
DISTANCE (AVG):	7.68E-01	7.81E-01	2%

As far as raw magnitudes are concerned, the B-C interaction system is definitely showing greater attractions. Furthermore within B-C, one can see that the mutant version of the protein exhibits greater interactions for both potentials. This is in contrast to the "second place" B-D system that favors wild-type interactions (again, for both potentials). Interestingly enough, although the B-D system is weaker overall, it nonetheless shows the most dramatic changes in potential magnitudes going from the wild-type to the mutant when compared against B-C's percent differences. As has been typical thus far (with the exception of the uncommon results for the A-C interactions) the distances have remained relatively constant and have not changed much between wild-type and mutant strains.

Turning one's attention to the fine parsing results, we can see that overall the distribution of atoms in space responsible for binding is relatively the same among wild-type and mutant. Overall residue participation was low, but slightly greater for the B-C system. Below are the image results for this portion of the study,

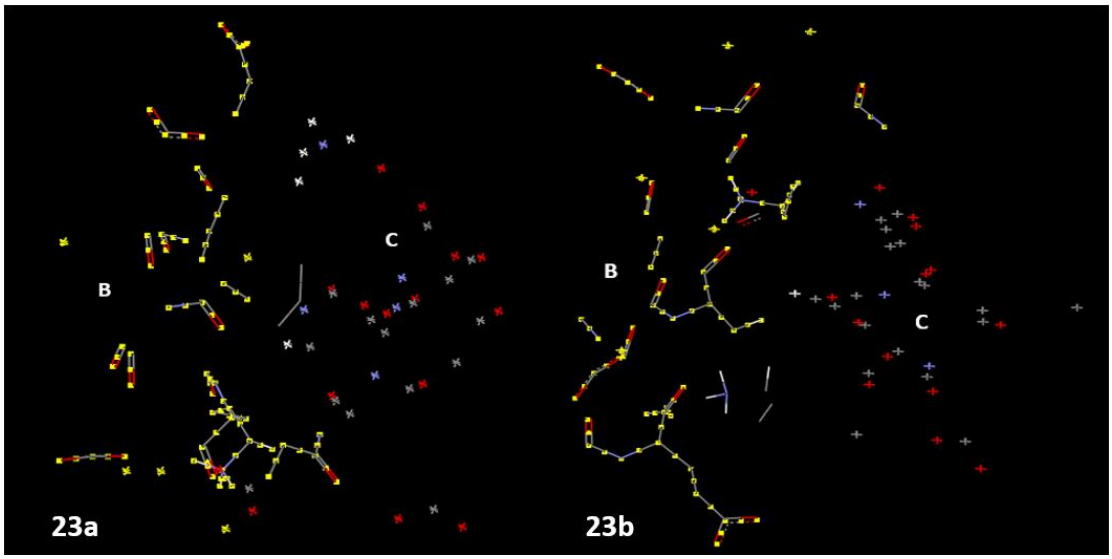


Fig. 23: P53 wild-type (23a) and mutant (23b) fine parsing, B-C

Wild-type: 343-354 (B chain), 343-354 (C chain)

Mutant: 342-354 (B chain), 343-354 (C chain)

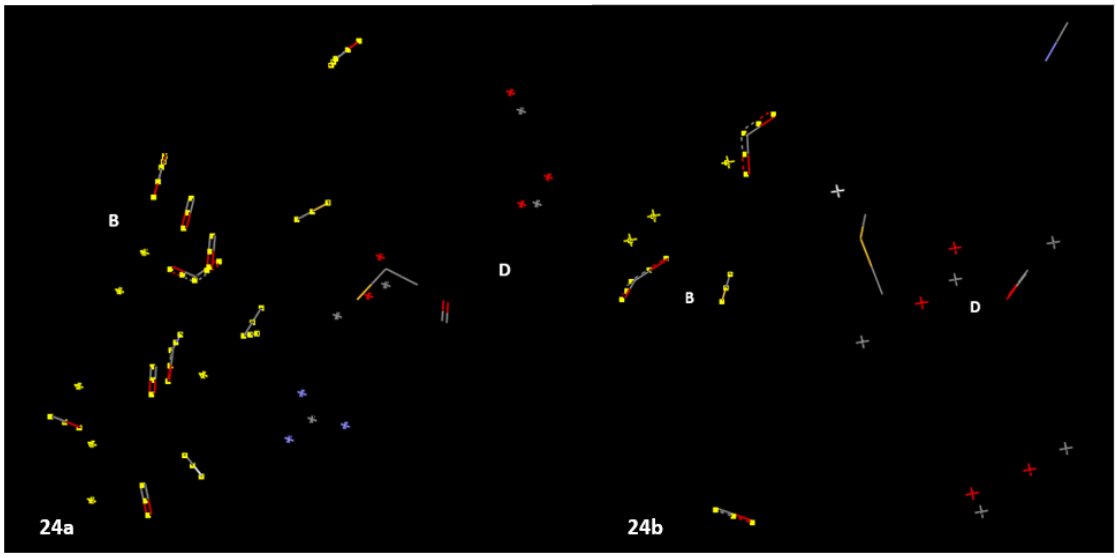


Fig. 24: P53 wild-type (24a) and mutant (24b) fine parsing, B-D

Wild-type: 712-916 (B chain), 336-343 (D chain)

Mutant: 336-344 (B chain), 336-344 (D chain)

Unlike in the A-X family, the mutated position (329 and 331) did not participate in binding in the B-X family of interactions.

Like its predecessors in the A-X module, the coarse parsing for all B-X interactions revealed that overall, the structure was generally well-preserved when going from the wild-type to the mutant. Especially observed with the B-C interactions, some individual residue's constituent moieties and atoms did acquire novel orientations/positions in the mutant strain, but apart from these slight details, there was little to suggest a major change in the structure concerned with the quatramer's intra-protein binding site. The images below detail these observations,

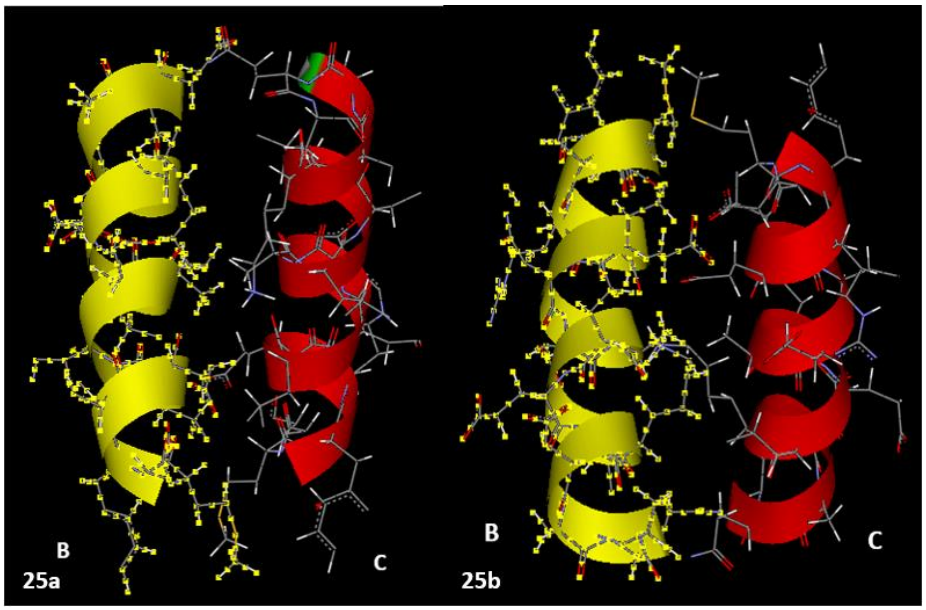


Fig. 25: P53 wild-type (25a) and mutant (25b) coarse parsing, B-C

Wild-type: 337-356 (B chain), 337-355 (C chain)

Mutant: 337-356 (B chain), 337-355 (C chain)

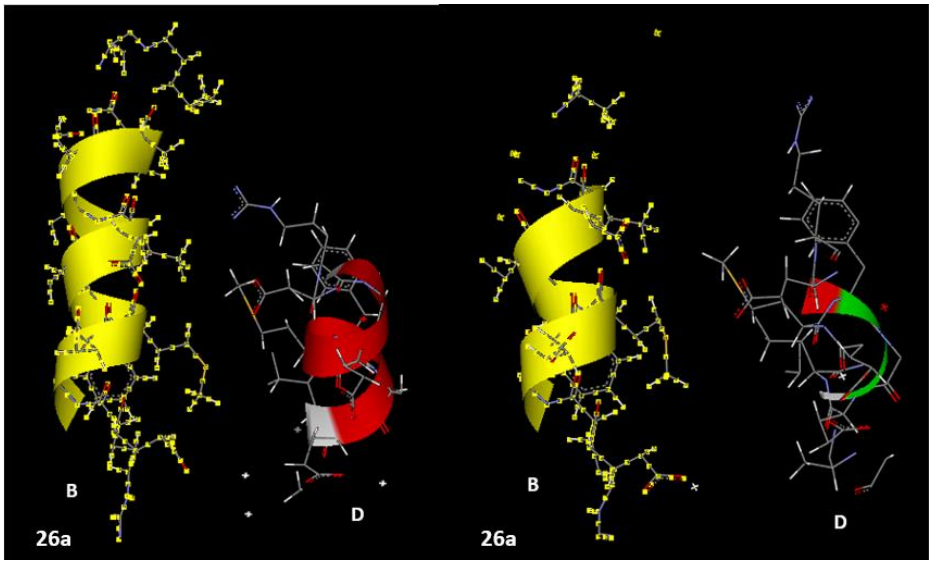


Fig. 26: P53 wild-type (26a) and mutant (26b), B-D

Wild-type: 336-351 (B chain), 335-351 (D chain)

Mutant: 336-351 (B chain), 336-351 (D chain)

Of a particularly interesting note was the slight reduction in alpha helical structure observed in the B-D system for both mutants' chains. Such a visual reduction further substantiates the fine parsing evidence that the wild-type stain is more stable and exhibits stronger intra-protein interactions. An item of note is the symmetry that the A-D chain interactions have with the B-C chain interactions. Such a similarity is taken from the perspective of the coarse parsing results for these two chains and is really not too surprising given the symmetry of the native protein.

Table 22: Common Results for p53 (B-C)

	WT	MUT	%
UL TOT:	-2.35E+00	-2.68E+00	12%
UCOUL TOT:	-4.59E+01	-5.42E+01	15%
DISTANCE (AVG):	7.27E-01	7.16E-01	2%
COMMON ATOM-ATOM PAIRS:	115		

Table 23: Common Results for p53 (B-D)

	WT	MUT	%
UL TOT:	-1.25E+00	-9.54E-01	24%
UCOUL TOT:	-8.08E+00	-8.78E+00	8%
DISTANCE (AVG):	7.94E-01	7.85E-01	1%
COMMON ATOM-ATOM PAIRS:	23		

The B-C system additionally showed that the mutant version of the protein was particularly stronger than the wild-type across both interaction criteria (Lennard-Jones and Coulombic). Thus the behavior of the common atom pairs were equal in their nature to the behavior of all the atom pairs from the overall analysis. As seen the foregoing analyses prior to the B-X family's, virtually every residue that participated in binding for the overall analysis was exactly represented in the analysis for common atom-atom interactions. There was some deviation from this observation for the wild-type B-D case which had a total of eleven residues present in binding interactions in the A chain alone. Out of these eleven residues, only five had constituent atoms that participated in common interactions with the mutant strain. The B-D

mutant strain by contrast had every single residue contribute to the common atom-atom interactions which again was typical of every system studied thus far. As before, distance data proved to conform more to typical values (between 7 to 9 dimensionless distance units).

Focusing on the uncommon results now, one can still see the B-C system reigning as the chain interaction pair with the most influential binding compared against the B-D system. These results are shown below,

Table 24: Uncommon Results for p53 (B-C)

	WT	MUT	%
UL TOT:	-1.39E+00	-4.47E+00	69%
UCOUL TOT:	-8.63E+00	-2.91E+01	70%
DISTANCE (AVG):	7.15E-01	6.61E-01	8%
UN-COMMON ATOM-ATOM PAIRS:	46	98	

Table 25: Uncommon Results for p53 (B-D)

	WT	MUT	%
UL TOT:	-1.23E+00	-2.95E-01	76%
UCOUL TOT:	-1.89E+01	-3.05E+00	84%
DISTANCE (AVG):	7.55E-01	7.74E-01	2%
UN-COMMON ATOM-ATOM PAIRS:	49	10	

The B-C system clearly shows superior interactions in both potential types within the mutant strain pertaining to a shift in interactions to both Lennard-Jones and Coulombic, just as it did in the common analysis thus indicating a stronger, better bound mutant. The B-D system by

contrast showed superior interactions in the wild-type strain with some considerable losses of both these Lennard-Jones and Coulombic potentials in the mutant thus yielding weaker B-D binding in the mutant.. The un-common analysis for the B-D interactions however proved to be slightly different in that the Coulombic potential for the wild-type was now considerably greater than its wild-type counterpart in the mutant strain (in contrast to the common analysis which showed the Coulombic interactions to be relatively equal). Thus the end result for the B-X chain interaction family is a small net overall change in the kinds of potentials present in those two chain interaction systems.

The final module for p53 that is now thus addressed is the C-D interaction system. As before, data for the overall analysis of the entire fine parsing output data file is shown below for this system,

Table 26: Fine Parsing Results for p53 (C-D)

	WT	MUT	%
UL TOT:	-5.92E+01	-5.31E+01	10%
UCOUL TOT:	-2.79E+02	-2.81E+02	1%
DISTANCE (AVG):	5.82E-01	5.91E-01	2%

As can be seen, the interaction parameters are relatively equal and thus do not change much when going from the wild-type to the mutant. Of interesting note however is the rather short distance (compared to the other analyses) that exists within both wild-type and mutant strains. So far only the analysis for the mutated position at position 329 (see Table 12) has yielded smaller values for inter-chain distances (5.47E-1 for the wild-type and 5.73E-1 for the mutant).

The fine parsing images indicated the presence of both mutated positions as participants in inter-chain binding between C and D for both the wild-type and mutant. This is reminiscent of the mutated positions likewise showing up in the fine parsing for the A-B system. The images showing these results are below,

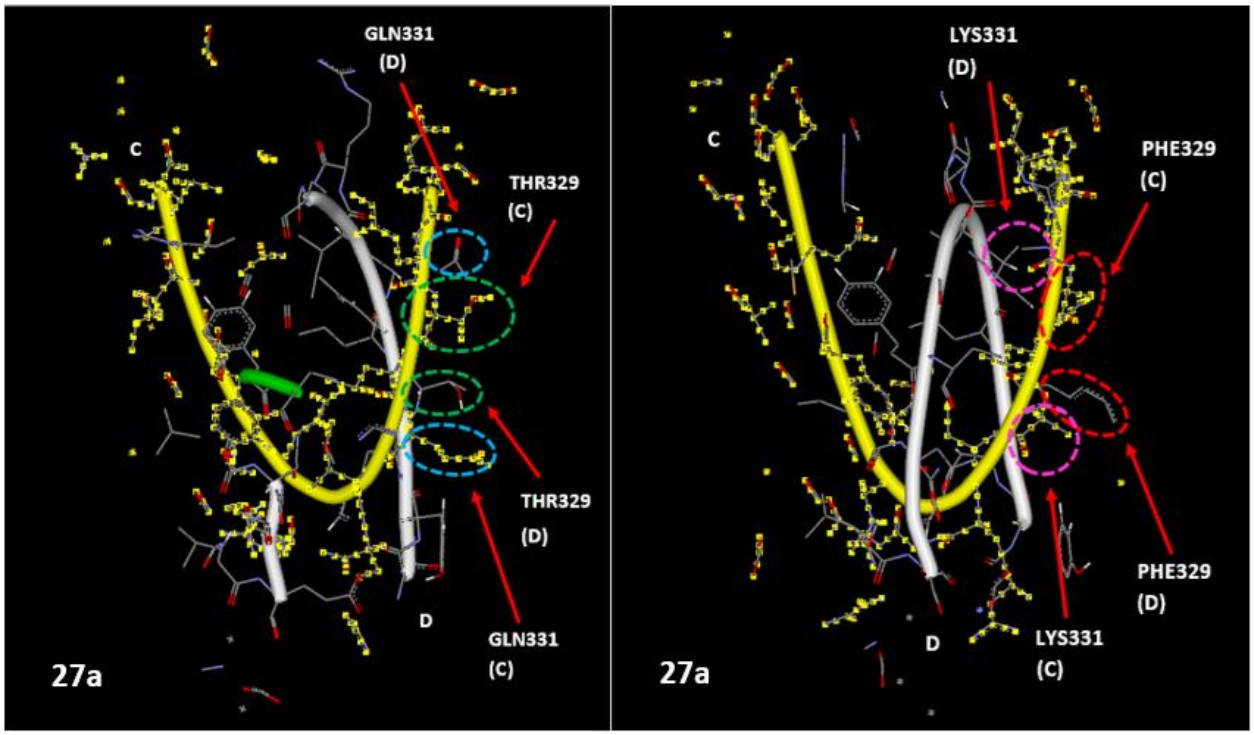


Fig. 27: p53 wild-type (27a) and mutant (27b) fine parsing, C-D

Wild-type: 326-352 (C chain), 326-352 (D chain)

Mutant: 326-352 (C chain), 326-356 (D chain)

As has been seen throughout this study, virtually every single residue had atoms that participated in fine parsing interactions. The only exceptions to this were the two mutated positions, 329 and 331, that are shown in the images above.

Continuing with the coarse parsing, we can see that like all of the preceding structures, the structure has been well-preserved and unchanged when going from the wild-type to the mutant strain. As pointed out before, some atoms and moieties managed to get slightly moved in orientation when introducing the mutations. As was observed with the A-X family, both mutations were present in the coarse parsing structures as well. The images below showcase these results,

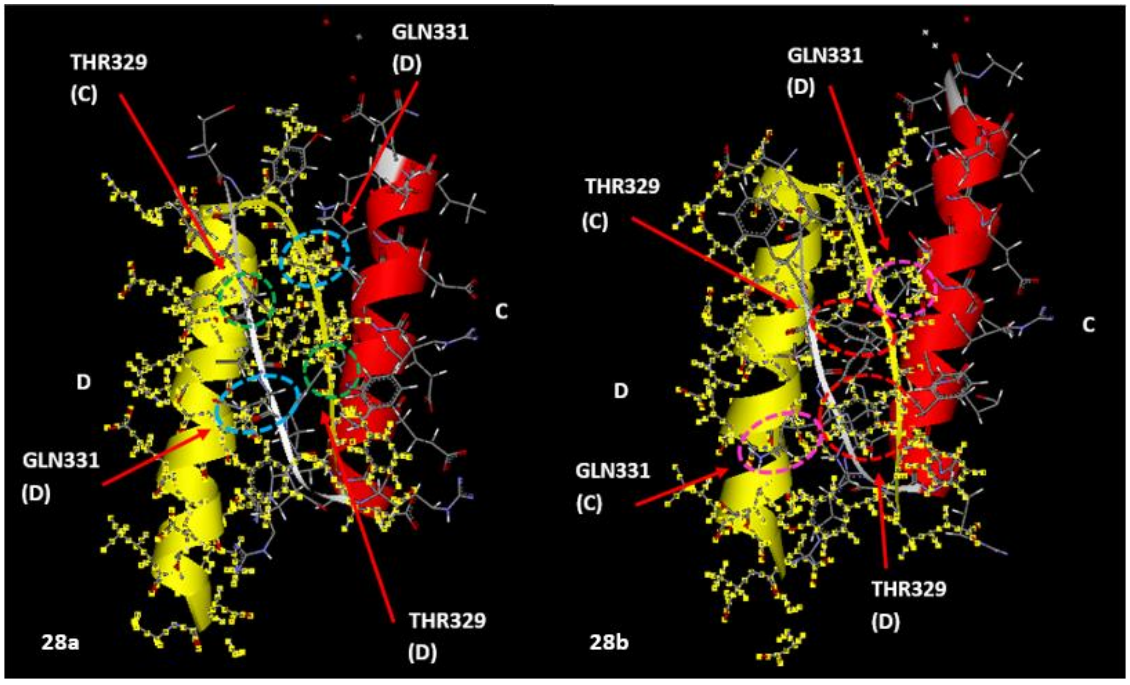


Fig. 28: p53 wild-type (28a) and mutant (28b) coarse parsing, C-D

Wild-type: 326-356 (C chain), 326-355 (D chain)

Mutant: 326-356 (C chain), 326-356 (D chain)

Turning attention to the common interactions, we can see that there were virtually no differences between the wild-type and mutant interaction pairs when going from one strain to

another. To put it another way, wild-type and mutant C-D interaction chain behaviors were the same.

Table 30: Common Results for p53 (C-D)

	WT	MUT	%
UL TOT:	-2.76E+01	-2.84E+01	3%
UCOUL TOT:	-1.79E+02	-1.83E+02	2%
DISTANCE (AVG):	6.09E-01	6.02E-01	1%
COMMON ATOM-ATOM PAIRS:	471		

When considering the un-common interactions, we see a relatively modest difference between the wild-type and the mutant. In all, the only significant observable difference was the wild-type losing Lennard-Jones interactions to the mutant by 22%, whereas the values for Coulombic interactions and inter-atomic distances remained very close to each other across each strain. The results are shown in the table below,

Table 30: Un-common Results for p53 (C-D)

	WT	MUT	%
UL TOT:	-3.16E+01	-2.47E+01	22%
UCOUL TOT:	-1.00E+02	-9.82E+01	2%
DISTANCE (AVG):	5.57E-01	5.79E-01	4%
UN-COMMON ATOM-ATOM PAIRS:	483	435	

Thus the wild-type remains as the better bound system for the C-D interaction pairs with Lennard-Jones interactions slightly going down in the mutated state.

Since the mutated positions showed up like in the C-D chain interactions (see Fig. 27), what follows now is an analysis that compares how the two residues behaved in terms of their constituent atoms/ potentials and distances. Below are the results of this inquiry,

Table 27: Mutant Position Comparison (329 & 331)

	WT	MUT	%
UL TOT:	-6.76E+00	-6.77E+00	0.1%
UCOUL TOT:	-6.55E+01	-4.78E+01	27%
DISTANCE (AVG):	6.33E-01	5.80E-01	8%

Remarkably enough, the results from all of the constituent atoms for both mutated positions behaved almost identically. Only the wild-type exhibited a considerable decrease in its Coulombic interactions (thus the wild-type was better bound) with the two values differing by 27%. As before, the distances remained very close to each other.

Focusing now on just the 329th residue position, we begin to see some deviation in potential behavior. These results are in the table below,

Table 28: Mutant Position Comparison (329)

	WT	MUT	%
UL TOT:	-3.22E+00	-4.50E+00	28%
UCOUL TOT:	-5.37E+01	-1.63E+01	70%
DISTANCE (AVG):	5.95E-01	4.99E-01	16%

Although the wild-type showed considerably stronger Coulombic interactions, the mutant strain ultimately displayed the most varied changes in the form of Lennard-Jones interactions

increasing, and the inter-atomic distances decreasing, all of which are hallmarks of improved and stronger interactions. The sharp decrease observed in the Coulombic interactions going from the wild-type to the mutant is consistent with what is known from the identity of the missense mutation covered here: the wild-type 329th residue is threonine which is a polar amino acid and it is replaced by a phenylalanine which is a non-polar amino acid. It thus makes sense that the polar influence should decrease given its replacement with something non-polar. In terms of distance, this was one of the few instances where the distance parameter changed noticeably, and for this case, it changed in favor of the mutant (the mutant is now closer among its chains).

Lastly, we now turn to the 331st mutated residue position. When considering only this position, the results indicated that the wild-type had superior interactions in both potentials and no changes to inter-atomic distances were observed. These results are shown below,

Table 29: Mutant Position Comparison (331)

	WT	MUT	%
UL TOT:	-3.54E+00	-2.27E+00	36%
UCOUL TOT:	-3.87E+01	-3.15E+01	19%
DISTANCE (AVG):	6.57E-01	6.57E-01	0%

Thus in all, one can see a wide variation of results for the different combinations of chains, even when they are within the same protein. In some cases the common interactions for the different strains are truly carbon copies of each, which is intuitive and makes sense, but in other scenarios, they are different enough to warrant further investigation using additional tools such as dynamic studies. For p53, it can thus be concluded that mutations have a competing (i.e. Lennard-Jones increasing but Coulombic decreasing at the same time) effect although weakening

the structure somewhat. The changes observed are not dramatic enough to prevent the tetramer from binding amongst its constituent chains, but they are sufficient to alter its binding ability to DNA.

5.3 SYSTEMS ANALYZED: HRAS

The next system analyzed was for another protein with strong associations with cancer (akin to p53): HRAS. The HRAS protein is made of a 21.45 kDa single chain ("Hras (human)," 2014) unlike the other protein systems analyzed thus far (which had two or more chains) so in this respect, its analysis approach was unique and required a novel consideration that was unnecessary before for the other systems (more on this later). Below is an image of HRAS wild-type that was employed for this study (accession number 4EFL) (Muraoka, Shima, Araki, & et al., 2012),

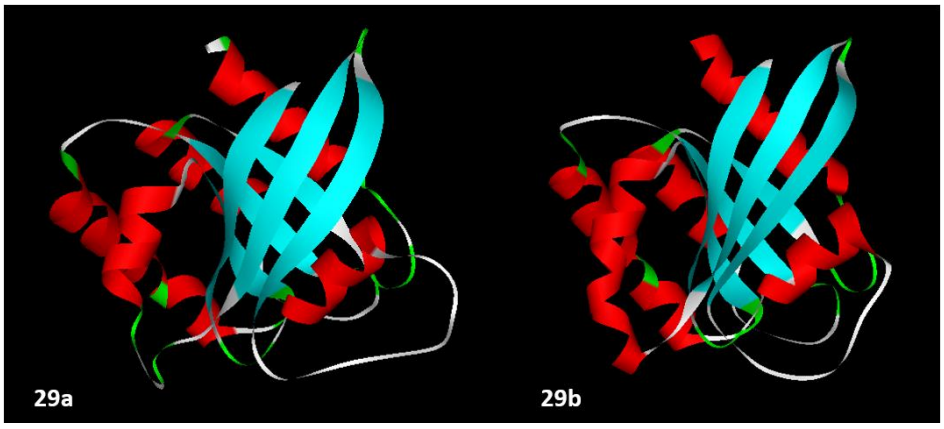


Fig. 29: HRAS wild-type (29a) and mutant (29b)

HRAS is a protein responsible for the regulation of cell division and cell growth. The regulation itself is triggered in response to growth factor stimulation (Muraoka et al., 2012).

Growth factors are a class of chemicals, namely, cytokines, hormones and proteins that positively affect cell growth and development (Muraoka et al., 2012). It is typically from an extracellular environment that HRAS interacts with these compounds and thus HRAS is usually associated with cell membranes due to the presence of an isoprenyl group on its C-terminus (Muraoka et al., 2012). Once the appropriate cell signals have been received, stimulation of signal transduction events by means of recruiting additional proteins such as c-Raf and PI 3-kinase within the cytoplasm direct the cell to continue growing and divide (a transduction event is the process by which proteins and secondary messengers relay signals from the outside of the cell to the cell nucleus) (Muraoka et al., 2012). Thus HRAS responds in an agonistic manner in response to the growth factors and is an early player in this particular cell signaling chain of command given its position at the front end of the process. The specific mechanism by which HRAS accomplishes its function in relaying the appropriate message in response to growth factor (or the absence thereof) is the enzymatic cleavage of GTP (guanosine triphosphate) to GDP (guanosine diphosphate) (Muraoka et al., 2012). When HRAS is bound to GDP, the protein is said to be off and it will not recruit downstream proteins to execute cell division. When HRAS is bound to GTP however, the protein is said to be on and instructions to carry out cell division and growth will be conveyed to the nucleus (Muraoka et al., 2012). Thus the process of turning HRAS "on" involves the enzymatically assisted hydrolysis of GTP to GDP specifically removing a phosphate group from the 5' carbon of the ribose sugar, and turning the protein "off" involves the re-phosphorylation of GDP to GTP accomplished with the help of pyruvate kinase, a phosphorylating enzyme (Muraoka et al., 2012).

With respect to the binding site of this protein, there are technically two, but only one is specifically involved in the pathology associated with aberrant HRAS forms. The image below details these binding sites as reported by Miyakawa, Morikawa, Takasu, & et al., 2013,

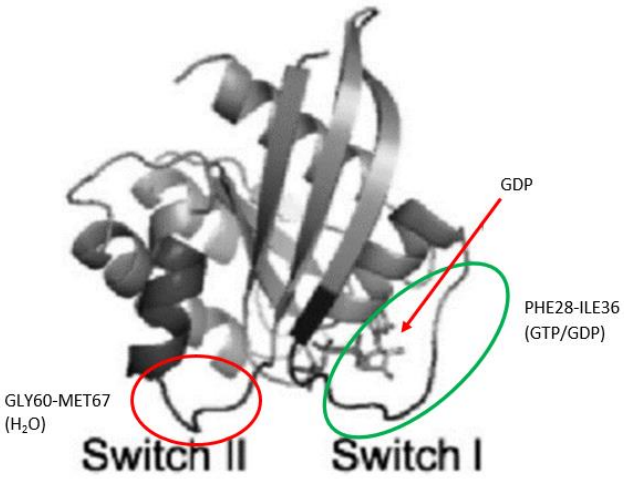


Fig. 30: HRAS binding sites, GDP complex (off)

With regards to the GTP/GDP dependency of the on/off status of HRAS, the loop in Figure 45 labeled as Switch I houses the GTP/GDP molecule mentioned earlier (Miyakawa et al., 2013). The second loop, called Switch II, in turn houses a molecule of water for the de-phosphorylation of GTP via hydrolysis (Miyakawa et al., 2013). When a mutation manifests itself in HRAS, the danger lies in how that mutation will affect the Switch I loop (Franken, Scheidig, Krengel, & et al., 1993). If the mutation should result in the "on" position (GTP bound state) remaining as a permanent state by preventing the normal enzymatic hydrolysis activity, then cancer may very well result due to the perpetual signal for the cell to grow and multiply (Franken et al., 1993). Thus wherein the danger with mutated p53 was in growth no longer being halted, the danger with mutated HRAS is that growth is now permitted excessively. Should HRAS be negatively

affected by a congenital mutation, then the result may be what is known as Costello syndrome (named after the New Zealand pediatrician who first reported this disorder to the Australian Paediatric Journal) (Franken et al., 1993). This disease is characterized by delayed development, mental retardation, cancerous (and non-cancerous) tumors, distinctive facial features, unusually flexible joints and heart abnormalities such as tachycardia (an extremely fast heartbeat) (Franken et al., 1993). Although the tumors are explained via the function of HRAS, researchers are as of yet still unsure as to how the other disease characteristics (especially mental retardation) manifest from aberrant HRAS production. The most common mutation resulting in Costello syndrome replaces the amino acid glycine at the twelfth position with serine (Franken et al., 1993). The mutation that was employed for this study replaced glycine at this same position but with aspartic acid instead (mutant accession number: 1AGP) (Franken et al., 1993). Both strains, wild-type and mutant, were in the "on," GTP bound state. The final location chosen was the 57th residue such that the first "sub" chain of HRAS would contain Switch I and the second "sub" chain would contain Switch II. Thus the first "sub" chain contained residues 1-57 and the second "sub" chain contained the residues 58-166 which marked the end of the HRAS chain. The image below shows this demarcation with the 57th residue hi-lighted in dark blue,

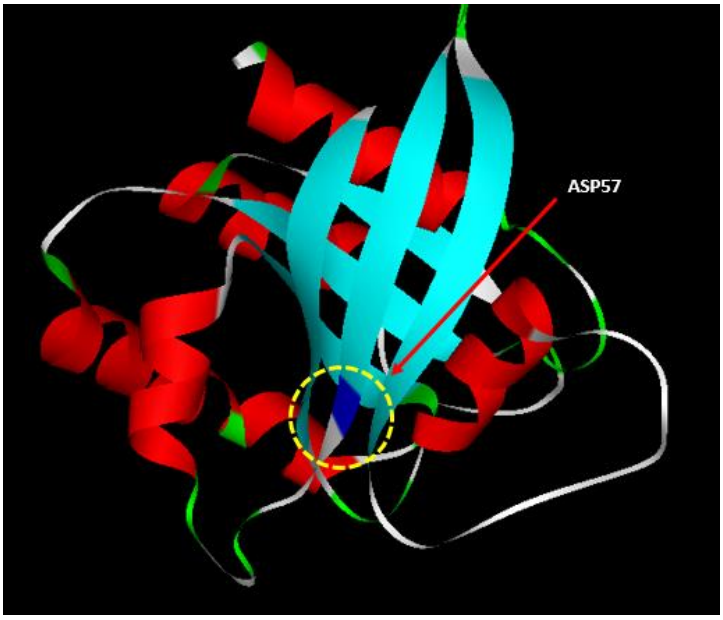


Fig. 31: HRAS splitting location

Table 30 shows the overall analysis for the Fine Parsing of the HRAS system as described above. We can see that the mutant strain had better binding interactions. Although both Lennard-Jones and Coulombic forces increased, the attractive Coulombic interactions increased the most when going from the wild-type to the mutant.

Table 30: Fine Parsing Results for HRAS

	WT	MUT	%
UL TOT:	-8.02E+01	-9.07E+01	12%
UCOUL TOT:	-3.81E+02	-4.98E+02	23%
DISTANCE (AVG):	6.09E-01	6.06E-01	0.5%

The first thing worth mentioning when considering the Fine Parsing images (Fig. 32) for the overall analysis of this system is that the specific arrangement of atoms in space changes

somewhat noticeably when going from the wild-type to the mutant strain. The other detail worthy of note is that the mutated residue position appears in the Fine Parsing Results for both strains and is therefore a direct participant in the interactions between Switch I and Switch II domains.

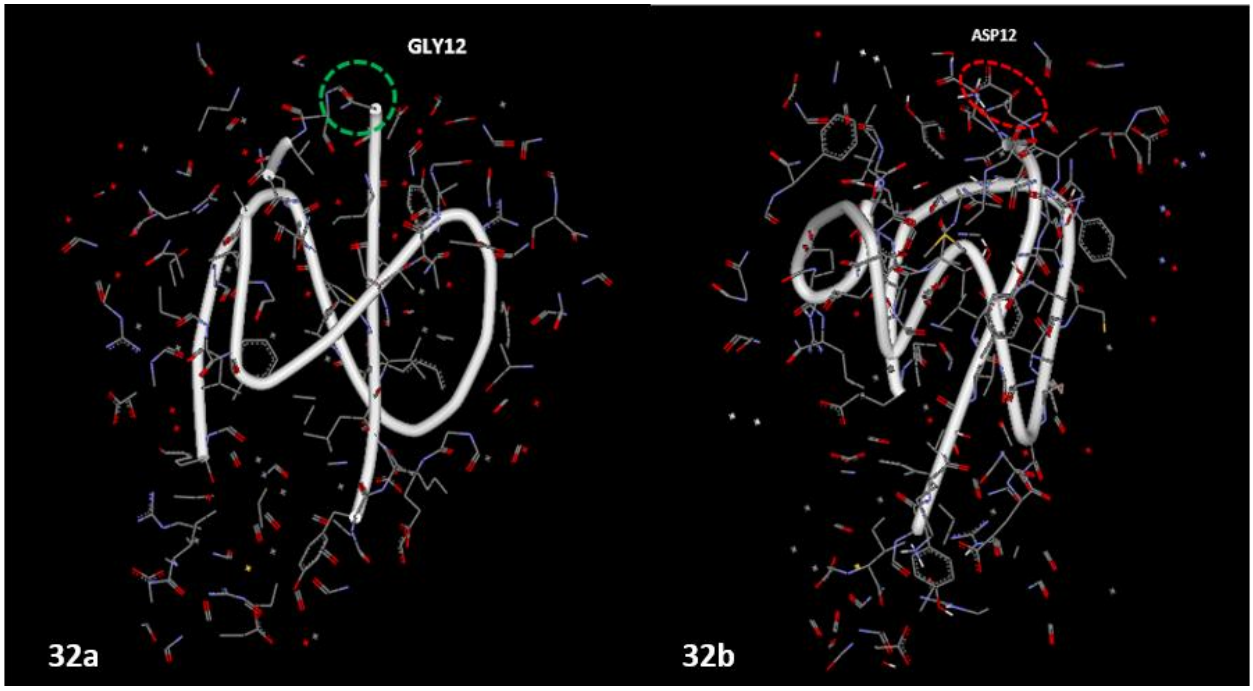


Fig. 32: HRAS wild-type (32a) and mutant (32b) Fine Parsing, sub-chains: 1-57 & 58-166

Wild-type: 1-57 (sub-chain 1-57), 58-165 (sub-chain 58-166)

Mutant: 1-57 (sub-chain 1-57), 58-165 (sub-chain 58-166)

It is interesting to note that despite the seemingly different arrangement of the atoms in the mutant compared against the wild-type, the data discussed in the foregoing paragraph nonetheless indicate that average inter-chain atomic distances remained virtually identical.

As can be seen from the coarse parsing analysis (Fig. 33), the Switch I loop is missing from the wild-type structure but is present in the mutant. Thus OpenContact's © results indicate that there indeed is a difference involving structure between the wild-type and the mutant to the conclusion that the Switch I loop in the mutant is more proximal to the Switch II loop than in the wild-type strain. This is consistent with the known pathological characteristics of mutated HRAS where the GTP in Switch I is negatively affected by somehow being shielded from de-phosphorylation hydrolysis (which is what should happen) thus leaving GTP permanently unchanged in the Switch I loop. This may lead to the observed rampant cellular growth and division associated with the mutant strain.

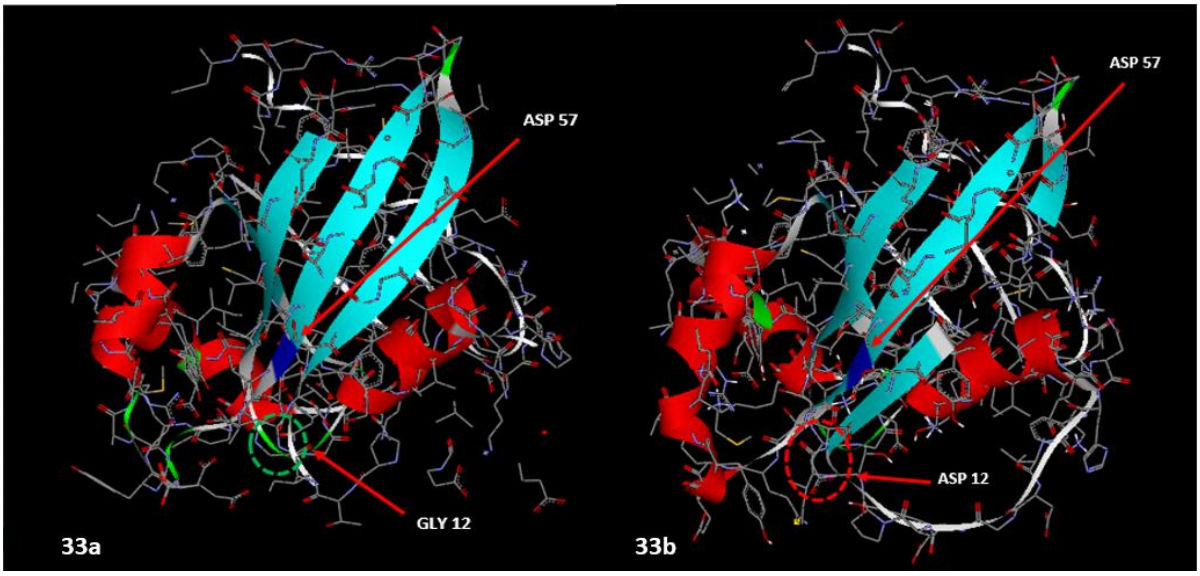


Fig. 33: HRAS wild-type (33a) and mutant (33b) coarse parsing, sub-chains: 1-57 & 58-166

Wild-type: 1-57 (sub-chain 1-57), 58-166 (sub-chain 58-166)

Mutant: 1-57 (sub-chain 1-57), 58-166 (sub-chain 58-166)

When considering the common interaction partners (Table 32), we can see that there is very little variation between the wild-type and mutant strains again.

Table 32: Common Results for HRAS

	WT	MUT	%
UL TOT:	-5.98E+01	-5.59E+01	7%
UCOUL TOT:	-3.46E+02	-3.50E+02	1%
DISTANCE (AVG):	6.08E-01	6.11E-01	1%
COMMON ATOM-ATOM PAIRS:	937		

When considering the un-common interactions (Table 33), we see a shift in interactions going in favor of the mutant, with a rather substantial increase, specifically in the Coulombic attractive interactions.

Table 33: Un-common Results for HRAS

	WT	MUT	%
UL TOT:	-2.04E+01	-3.48E+01	41%
UCOUL TOT:	-3.49E+01	-1.48E+02	76%
DISTANCE (AVG):	6.17E-01	5.99E-01	3%
UN-COMMON ATOM-ATOM PAIRS:	278	574	

Thus when comparing the uncommon atom-atom interaction pairs to the common interaction pairs, it can be seen that the changes effected to the mutant (of most importance to the Switch I loop) entirely come from the uncommon interactions. Furthermore, these uncommon interactions are considerably stronger for the mutant thus indicating that whatever hindrance is imposed upon GTP in the Switch I loop stems from an inherently in-flexible loop that does not permit the catalytic cleavage that this enzyme is supposed to execute.

An interesting result that had not been observed thus far was that although out of all the wild-type's residues, each participated in common binding (except the position that was to be mutated of course), the mutant itself had five extra residues that did not participate at all in common interactions with the wild-type strain. These extra residues in the mutant that were foreign to the wild-type's binding were residues 27-30. The wild-type strain thus exhibited 48 residue interactions and the mutant 53.

Upon close inspection and comparison of the mutated residue in the wild-type and mutant, one can see that they are both remarkably similar. The only real difference is a modest gain in Lennard-Jones interactions by the mutant which is curious in that the aspartic acid that defines the mutant is charged, more so than the glycine in the wild-type. Thus one would expect an increase in Coulombic interactions, but reality defied this expectation. The results are presented in the table below,

Table 31: Mutant Position Comparison (12)

	WT	MUT	%
UL TOT:	-1.22E+00	-1.67E+00	27%
UCOUL TOT:	-8.80E+00	-8.73E+00	1%
DISTANCE (AVG):	2.21E+01	2.02E+01	9%

As has become a hallmark of these systems, the distance remained unchanged is the closest by far among the specific cases wherein a mutated residue position participated in binding.

5.4 SYSTEMS ANALYZED: DHFR

The next system protein studied was the 21.45 kDa, enzyme dihydrofolate reductase, or DHFR ("DHFR (human)," 2014). Like HRAS, this protein was also a one chain monomeric system. The wild-type structure that was employed for this study was of the accession number 1DRF (Oefner, D'Arcy, & Winkler, 1988).

DHFR is an enzyme produced in the liver that is responsible for the reduction of dihydrofolic acid to tetrahydrofolic acid using the cofactor nicotinamide adenine dinucleotide phosphate (NADPH) as an electron donor (see image below),

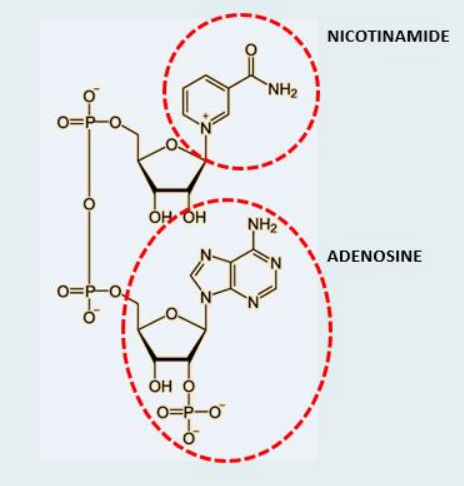


Fig. 34: NADPH with DHFR-relevant moieties

Tetrahydrofolic acid is an important precursor in the in-vivo bio-synthesis of purines and pyrimidines which in turn serve as the building blocks of DNA and RNA (Oefner, D'Arcy, & Winkler, 1988). DHFR thus serves as a catalyst for the important building blocks of genetic material and it is therefore found in every dividing cell, eukaryotic or otherwise. The mechanism

of enzymatic activity was stated by Oefner et al. as being the transfer of a hydride from NADPH to dihydrofolate with an accompanying protonation to produce tetrahydrofolate. As will be explored in more detail shortly, an important location on this protein known as the Met20 loop helps to stabilize the nicotinamide ring of the NADPH (see Figure 52) to promote the transfer of the hydride from the NADPH to the dihydrofolate (Ulrich, Akutsu, Doreleijers, & et al., 2014) .

DHFR is generally divided into two major sub-domains: the adenosine-binding sub-domain and the loop sub-domain (Cario , Smith, Blom, & et al., 2011). The adenosine-binding subdomain is the larger of the two and it binds the adenosine moiety of NADPH (see Figure 52) (Cario et al., 2011). The loop sub-domain contains three loops, the Met-20 loop (mentioned in the foregoing paragraph), the F-G loop and the G-H loop (Ulrich et al., 2014). In between these two sub-domains lies a long groove structured by a large beta sheet and an alpha helix which is the active site where folate and NADPH bind together (Ulrich et al., 2014). The overall location of these structures and the active site is shown in the following image as reported by Ulrich et al.,

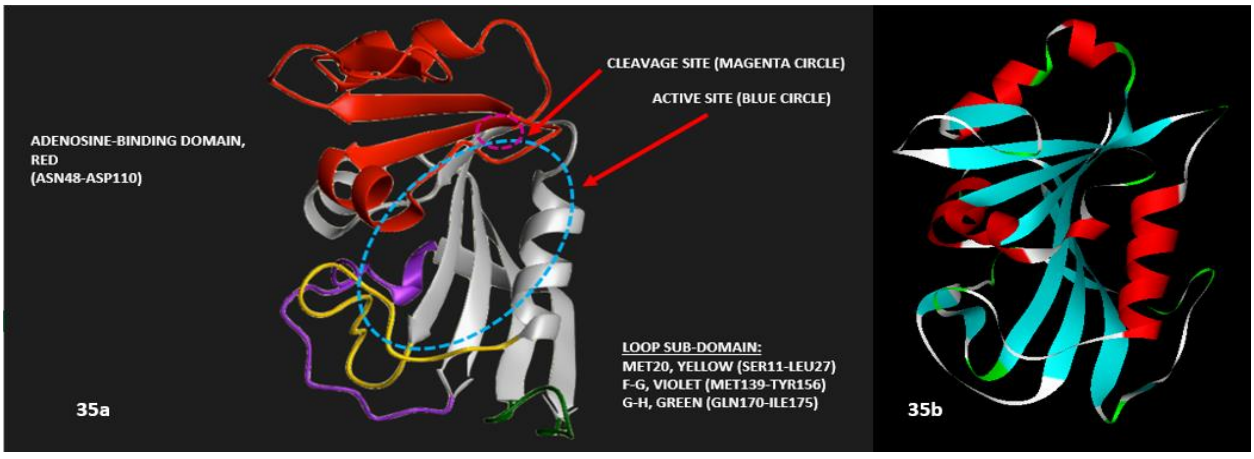


Fig. 35: DHFR wild-type with ligand binding domains (35a) and mutant (35b).

Cleavage site: VAL112. Wild-type accession number: 1DRF, mutant accession number: 3EIG.

As will be expounded upon shortly, the binding behavior of this protein that can affect human health is the binding of NADPH to DHFR (Oefner et al., 1988).

Should mutations occur near or on the loop sub-domain, DHFR may have problems adequately binding to NADPH, specifically, the nicotinamide moiety (Volpato, Yachnin, Blanchet, & et al., 2009). As reported by Ulrich et al., the Met 20 loop regulates NADPH binding by being flexible and allowing the cofactor access to its binding site. Thus mutations that affect the loop sub-domain, specifically the Met 20 loop, may have negative consequences in DHFR even binding to its NADPH ligand and carrying out its function. If such a congenital mutation were to manifest itself, a disorder known as megaoloblastic anemia can occur (Volpato et al., 2009). In this disorder, a deficiency of tetrahydrofolate (the product of DHFR and dihydrofolate) results in inadequate DNA synthesis for red blood cell production (Volpato et al., 2009). When DNA synthesis is impaired, the cell cycle cannot continue beyond the G2 growth stage to the M (mitosis) stage. Thus a common hematological trademark of megaloblastic anemia are (as the name "megaloblastic" might imply) abnormally large and dysfunctional red blood cells in the bone marrow. Like many other kinds of anemia, symptoms of this variety of the disease are fatigue, weakness, difficulty concentrating and general malaise (Volpato et al., 2009). For this study, a mutant strain was analyzed with two mutations: F31R and Q35E These mutations were both located on the alpha helix superior to the loop sub-domain. The mutant's accession number was 3EIG (Volpato et al., 2009). Since DHFR is a monomeric protein, it had to be cleaved in order to feed it to OpenContact©. The location of thus breaking the pdb file was position 112 and it was so chosen so that the two sub-domain NADPH binding regions would be separate (see Figure 35).

The results for the overall Fine Parsing analysis of this system showed that the mutant had gained Lennard-Jones interactions, although the gain was relatively modest. The table below details these results,

Table 32: Fine Parsing Results for DHFR

	WT	MUT	%
UL TOT:	-1.31E+02	-1.53E+02	14%
UCOUL TOT:	-4.56E+02	-4.77E+02	4%
DISTANCE (AVG):	5.87E-01	5.79E-01	1%

The pictorial fine parsing results for DHFR show that overall three-dimensional binding atom distribution remained relatively constant between the two strains. Of note was that residues of the Met 20 loop were present in the fine parsing files for both strains indicating that it did play a direct, significant role in intra-protein interactions. The Met 20 loop region is encircled in yellow in the Fig. 36.

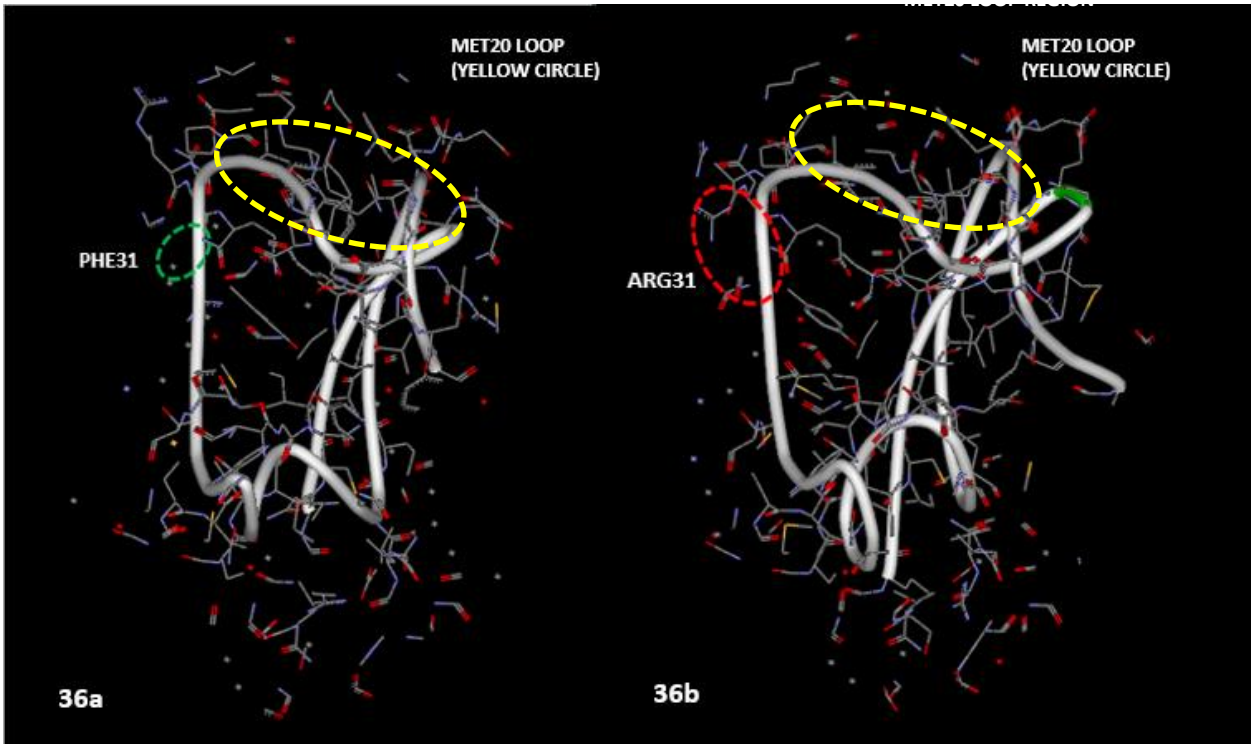


Fig. 36: DHFR wild-type (36a) and mutant (36b) Fine Parsing, sub-chains: 1-112 & 113-186

Wild-type: 2-112 (sub-chain 1-112), 113-184 (sub-chain 113-186)

Mutant: 2-112 (sub-chain 1-112), 113-184 (sub-chain 113-186)

Also 31st residue position that was ultimately mutated, played a role in the wild-type and mutant interactions. It was also observed that every residue that participated in fine parsing interactions had constituent atom-atom pairs that were common to both strains.

When considering the coarse parsing structures, we can see that the mutant strain acquired more structure in that it now has an additional helix that did not play a role in the wild-type. Apart from this detail, the remainder of the structure was identical. The images are below,

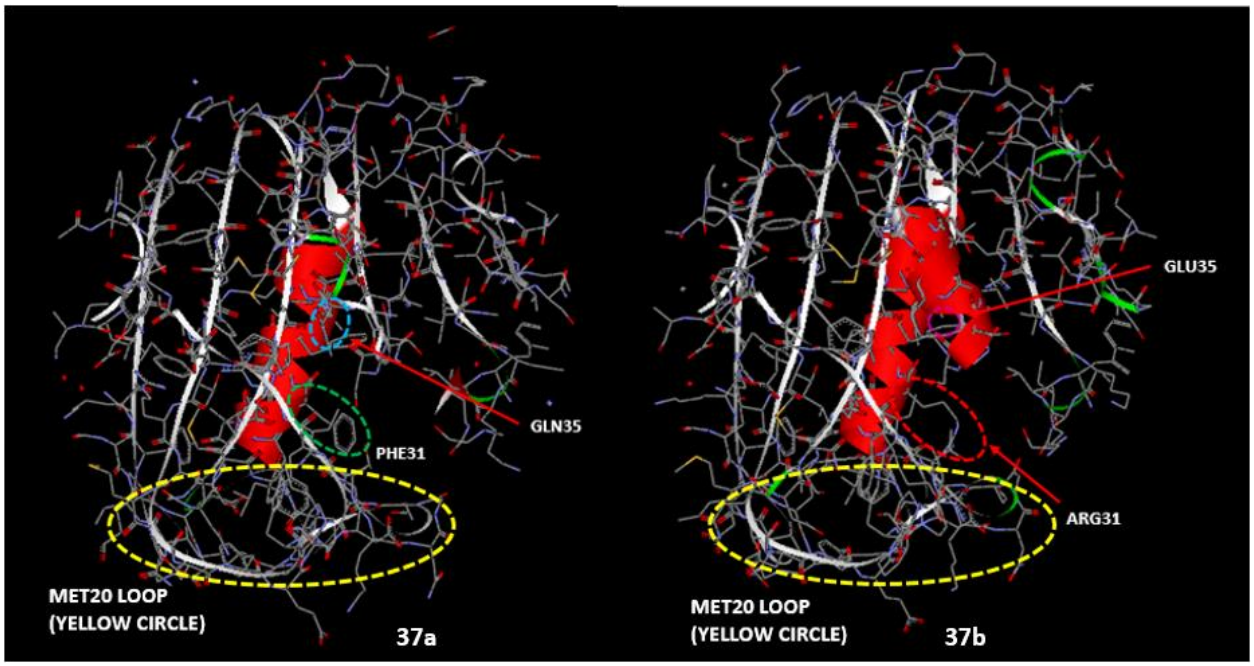


Fig. 37: DHFR wild-type (37a) and mutant (37b) coarse parsing, sub-chains: 1-112 & 113-186

Wild-type: 1-112 (sub-chain 1-112), 113-186 (sub-chain 113-186)

Mutant: 1-112 (sub-chain 1-112), 113-186 (sub-chain 113-186)

Again, as in the fine parsing image results, the 31st residue showed up as a structural participant in both strains. The 35th position also showed up here, though not in the fine parsing results.

The common results showed practically no variation between the wild-type and the mutant. These results are in the table below,

Table 36: DHFR Common Results

	WT	MUT	%
UL TOT:	-1.02E+02	-1.06E+02	4%
UCOUL TOT:	-4.79E+02	-4.89E+02	2%
DISTANCE (AVG):	5.97E-01	5.90E-01	1%
COMMON ATOM-ATOM PAIRS:	1427		

The results for the un-common interactions indicated a curious balance between Lennard-Jones interactions and Coulombic repulsion. Both strains have an attractive Lennard-Jones criterion and repulsive Coulombic interactions. Of particular note was the dominant shift in Lennard-Jones interactions that the mutant strain exhibited compared to the wild-type (the wild-type had a Lennard-Jones to Coulombic ratio of approximately 1.2 compared to the mutant which had a Lennard-Jones to Coulombic ratio of around 3.9).

Table 37: DHFR Un-common Results

	WT	MUT	%
UL TOT:	-2.84E+01	-4.62E+01	39%
UCOUL TOT:	2.29E+01	1.20E+01	48%
DISTANCE (AVG):	5.36E-01	5.44E-01	1%
UN-COMMON ATOM-ATOM PAIRS:	281	447	

The results for comparing the behavior of both the wild-type and mutant residue (position 31) yielded results that indicated that the mutant strain of the protein had superior binding potentials compared to the wild-type. The table below summarizes these results,

Table 34: DHFR mutant position (31)

	WT	MUT	%
UL TOT:	-9.35E-03	-2.60E-02	64%
UCOUL TOT:	-1.39E+00	-2.18E+00	36%
DISTANCE (AVG):	8.97E-01	7.75E-01	14%

As was observed with the overall analysis, the mutant residue showed the steepest increase in Lennard-Jones potentials which seems counter-intuitive. The expectation of the mutant's charged arginine amino acid would be that it should have a greater increase in Coulombic interactions than the wild-type.

5.5 SYSTEMS ANALYZED: HUNTINGTIN

The final protein analyzed in this study was the huntingtin protein (HTT). It is a trimer protein (a three chain system) with each monomer having a molecular weight of 347.60 kDa for a total of 1042.8 kDa for the whole, biologically active protein ("Huntingtin (human)," 2014). As the name suggest, this protein is implicated in Huntington's disease, a neurodegenerative disorder. Due to the extreme difficulty in crystallizing this protein, the following structure below for the wild-type strain of what was studied is only of the structure of the N-terminal region (accession number 3IO4),

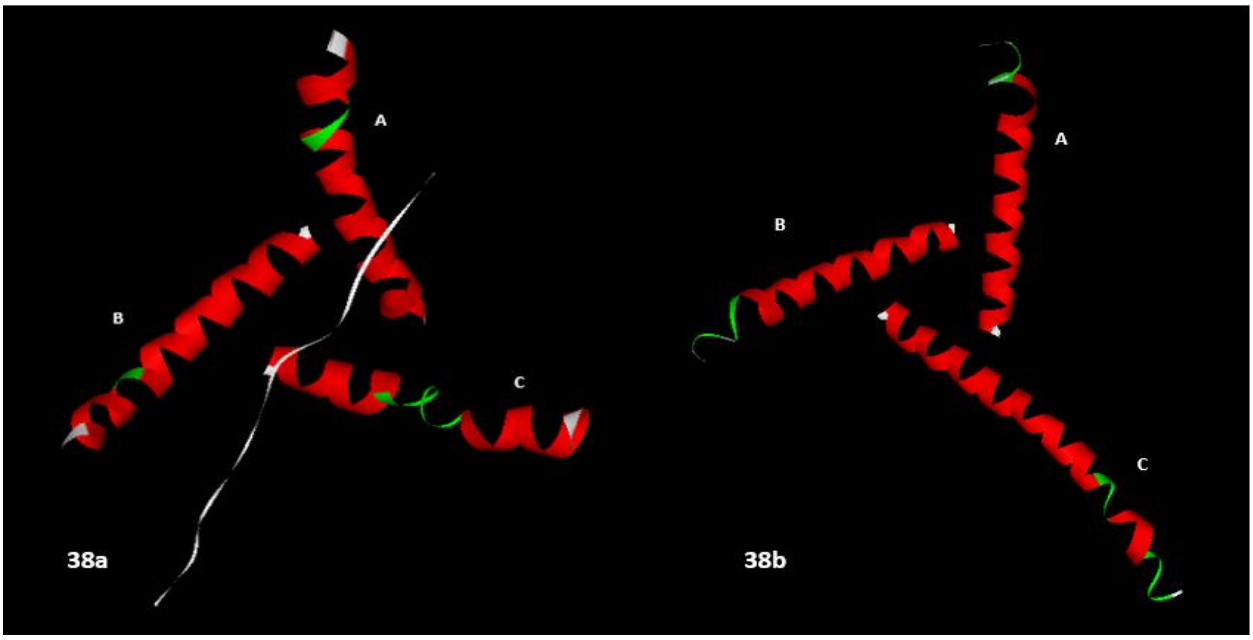


Fig. 38: HTT wild type N-terminus (38a) and mutant (38b)

Wild type accession number: 3IO4

Mutant accession number: 4FE8

It is important to note that the pdb file corresponding to the above image was of the Huntingtin protein's N-terminus stabilized by a maltose binding protein to help retain structural stability during crystallization (Kim, Chelliah, Kim, Otwinowski, & Bezprozvanny, 2009). This maltose binding protein was manually omitted for the study however and only this segment of the huntingtin protein was subjected to analysis. As with p53, this system involved considering the interactions from the different combinations of chain interaction partners (i.e. A-B, A-C, etc…).

Huntingtin is an elusive protein in that its exact role in vivo is not clearly understood (Kim et al., 2009). Experiments with transgenic mice where the gene that encodes for huntingtin has been deleted (knockout) have proven lethal to the rodents (Kim et al., 2009). This protein is highly expressed in neurons and testes of both humans and rodents and interestingly, it has no sequence homology with any other protein (Kim et al., 2009). Huntingtin has also been shown to upregulate the expression of the nerve factor Brain Derived Neurotrophic Factor (BDNF) at the transcription level, but the exact mechanism of how this is accomplished is not known either (Kim et al., 2009). From various experiments in electron microscopy and immunohistochemistry, it has been found that Huntingtin is primarily associated with the intra-cellular machinery of vesicles and microtubules (Kim et al., 2009).

Although Huntingtin is known to interact with many proteins in the human body (most famously the Huntingtin Interacting Protein or HIP), due to its presently unknown mechanism of

function, there is really no active site to presently report of for this protein (although the protein itself is ubiquitously found in the cytoplasm of cells) (Kim et al., 2009). Thus the perspective of this study will instead focus on a unique aspect of mutated Huntingtin's pathology: a longer than normal polyglutamine segment (known as a polyglutamine expansion) that is associated with Huntingtin aggregates in infected systems (Kim, 2013). Huntingtin's disease occurs when the polyglutamine expansion exceeds 36 glutamines near the amino terminus although the cellular mechanisms that link this expansion to disease manifestation is still under investigation (Kim, 2013). The following image shows the location of the polyglutamine region in the mutant strain that was studied (accession number 4FE8) (Kim, 2013). This mutant strain had thirty-six glutamine residues,

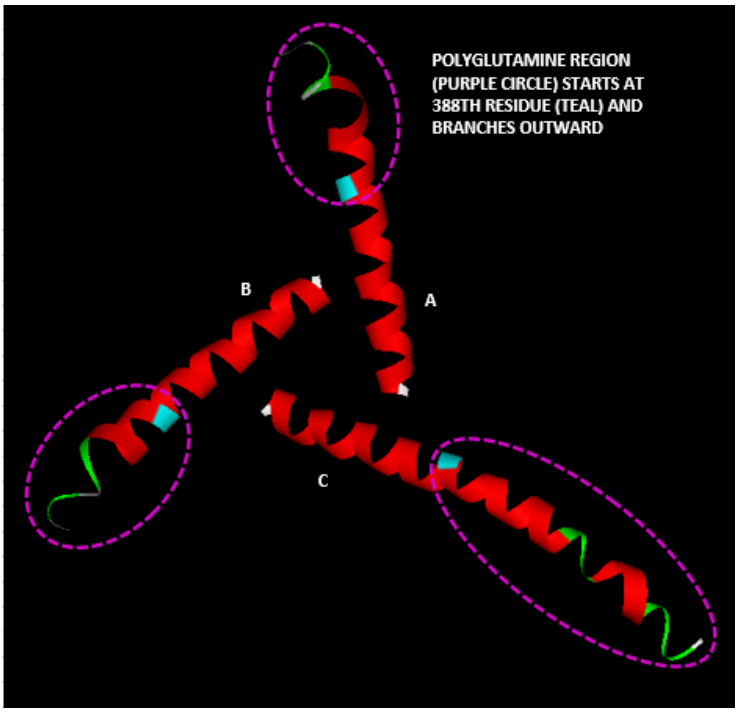


Fig. 39: Huntingtin mutant (accession number: 4FE8)

The glutamine expansion region starts with the 388th residue (Kim, 2013). It is important to note that not all thirty-six glutamines in the polyglutamine region were captured by the crystallographic experiment that yielded the above structure. As one can envision from the C chain however (which presented the most complete polyglutamine region), the polyglutamine region would simply extend all three chains in the figure's triangular configuration outward. How this extension would affect inter-chain binding (if it affects it at all) is the subject of this study. With respect to pathology, most evidence is consistent with the hypothesis that the expanded protein acquires a "toxic gain of function (Kim, 2013)." Apart from the aggregates, other toxic properties incumbent on expanded Huntingtin have been observed to be negative effects on gene transcription, induction of apoptosis, and disruption of key neuronal functions such as axonal transport, synaptic transmission, and Ca^{+2} signaling (Kim, 2013). As stated by Kim et al., 2009, "many of the proposed mechanisms suggest that expanded Huntingtin is involved in pathological interactions with other signaling proteins in cells, leading to neuronal dysfunction and death." The mutant strain studied was a form of Huntingtin with thirty-six glutamines: pdb accession number: 4FE8.

The first combination that will be considered in the overall analysis is the A-X family of chains. As can be seen from Tables 38 and 39, there were some substantial shifts in the nature of the potentials that dominated between the wild-type and the mutant. In general, the most dramatic changes were observed to take place between the A-C chain interactions.

Table 38: Fine Parsing Results for Huntingtin (A-B)

	WT	MUT	%
UL TOT:	-7.83E+00	-3.31E+00	58%
UCOUL TOT:	-1.32E+01	-1.24E+01	6%
DISTANCE (AVG):	5.57E-01	6.42E-01	13%

Table 39: Fine Parsing Results for Huntingtin (A-C)

	WT	MUT	%
UL TOT:	-8.20E+00	-3.05E+00	63%
UCOUL TOT:	-8.86E+00	-1.11E+01	20%
DISTANCE (AVG):	5.68E-01	6.45E-01	12%

In the A-B interactions there was an overall reduction in the attractive forces in going from wild-type to mutant whereas in A-C, there were compensating effects of decreases in Lennard-Jones attractions and increases in Coulombic attractions. An interesting item of note is that the presence of a polyglutamine region in the mutant apparently creates more distance between the chains for both strains (the mutant average distance is longer in both). In general though, the similarity between A-B and A-C is not too surprising given the near equilateral triangular geometry that Huntingtin's N-terminus (see Figure 58).

Considering the pictorial representation for the fine parsing results (Fig. 40), we can see that the polyglutamine region participated in binding effects for the A-B wild-type interactions, but not for for the A-B mutant despite its expanded role in the mutant.

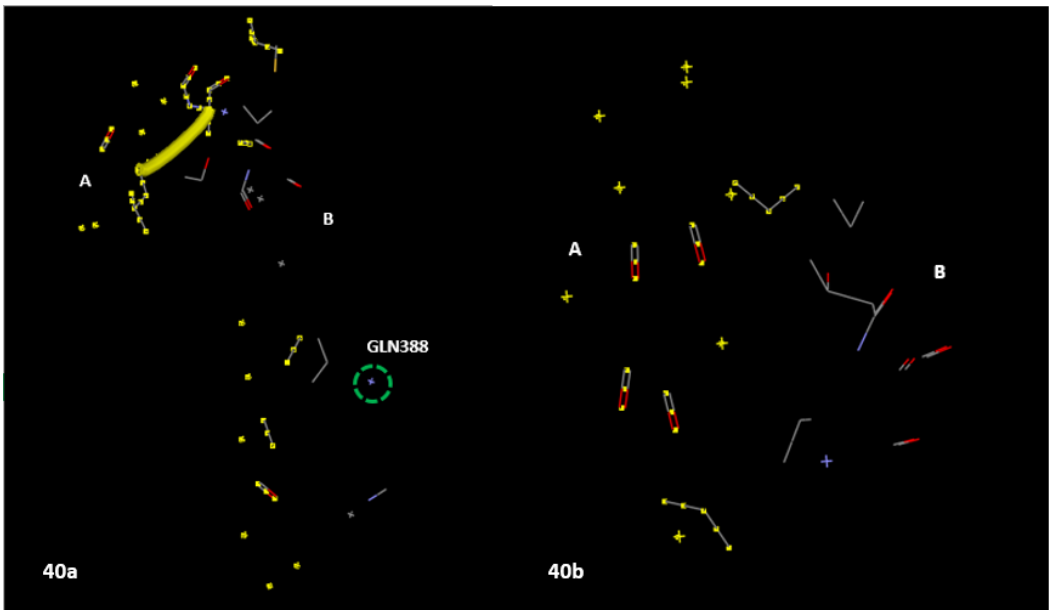


Fig. 40: Huntingtin wild-type (40a) and mutant (40b) Fine Parsing, A-B

Wild-type: 373-413 (A chain), 371-391 (B chain)

Mutant: 372-384 (A chain), 371-377 (B chain)

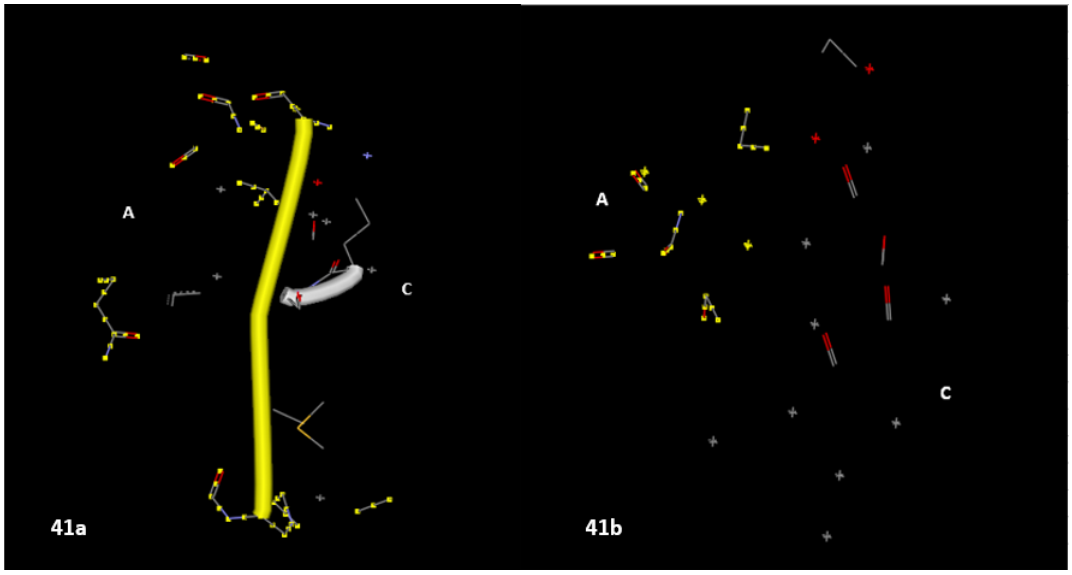


Fig. 41: Huntingtin wild-type (41a) and mutant (41b) Fine Parsing, A-C

Wild-type: 371-412 (A chain), 371-384 (C chain)

Mutant: 371-377 (A chain), 372-384 (C chain)

No presence of the glutamine region was found in the any of the A-C interactions, wild-type or mutant (Fig. 41).

The coarse parsing results showed that the wild-type of both A-B and A-C chains had more secondary structure than the mutant which is consistent with the increased atom-atom separation distances (Figs. 42 and 43). The A-B and A-C chain wild-types furthermore had the polyglutamine region participating in structure. Only the A-C system's mutant had the polyglutamine region however. These results are presented in the images below,

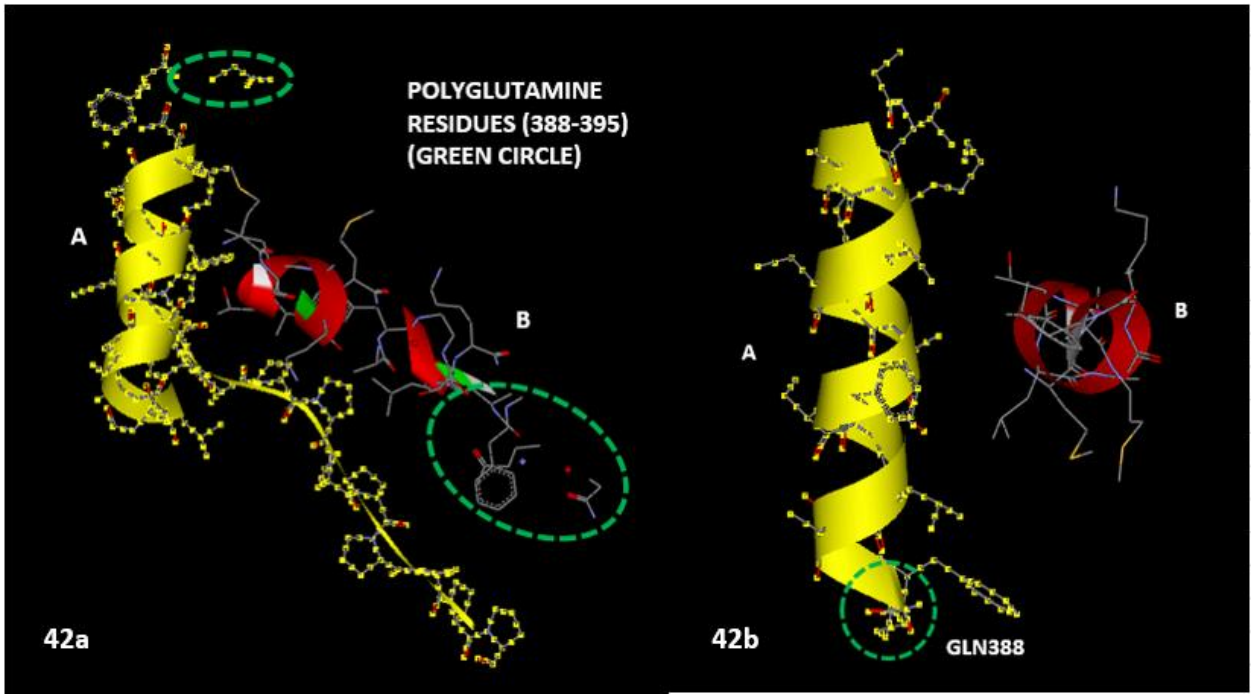


Fig. 42: Huntingtin wild-type (42a) and mutant (42b) coarse parsing, A-B

Wild-type: 372-415 (A chain), 371-395 (B chain)

Mutant: 372-388 (A chain), 371-381 (B chain)

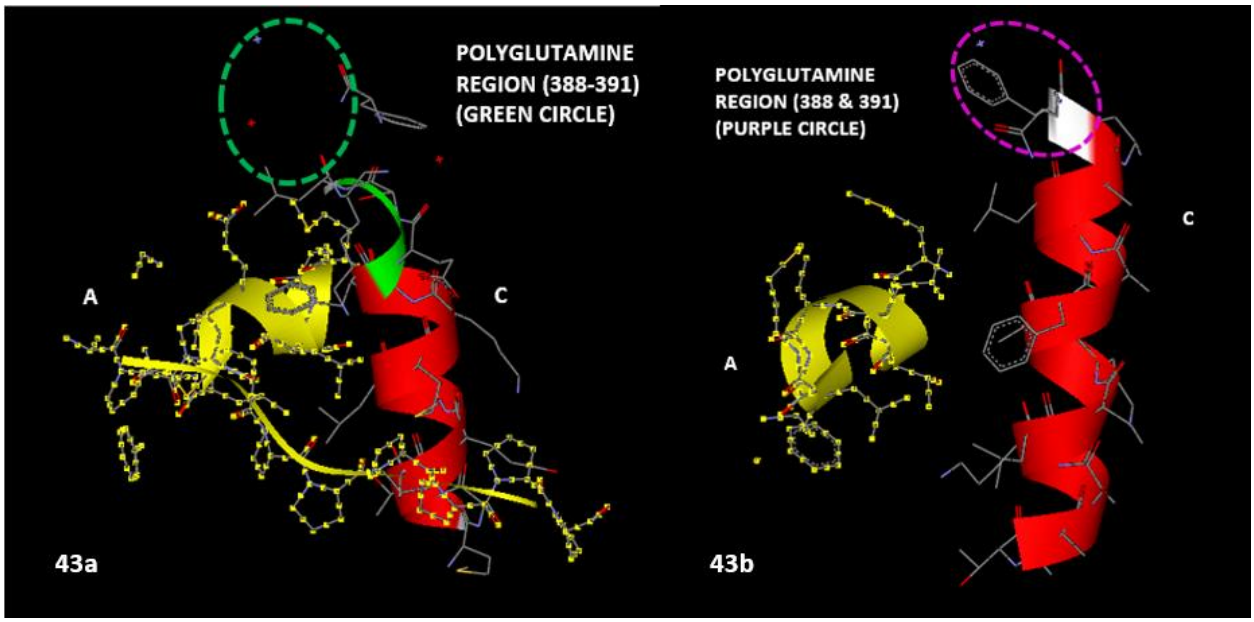


Fig. 43: Huntingtin wild-type (43a) and mutant (43b) coarse parsing, A-C

Wild-type: 371-413 (A chain), 371-391 (C chain)

Mutant: 371-381 (A chain), 372-391 (C chain)

Looking at the common atom-atom interaction results from the Fine Parsing data (Tables 40 and 41), we see, first of all, that the A-B interaction system was not truly affected in any meaningful way. For the A-C chain system however, we see that the mutant results were significantly different from the wild-type even for the same of common atom-atom pair interactions.

Table 40: Common Results for Huntingtin (A-B)

	WT	MUT	%
UL TOT:	-8.20E-01	-7.77E-01	5%
UCOUL TOT:	-9.34E+00	-9.05E+00	3%
DISTANCE (AVG):	6.94E-01	6.95E-01	0.2%
COMMON ATOM-ATOM PAIRS:	23		

Table 41: Common Results for Huntingtin (A-C)

	WT	MUT	%
UL TOT:	-6.77E-02	-1.43E-01	53%
UCOUL TOT:	-5.37E+00	-6.77E+00	21%
DISTANCE (AVG):	8.22E-01	7.54E-01	8%
COMMON ATOM-ATOM PAIRS:	14		

Upon studying the un-common results for the A-X family, we first of all see that both inter-chain interaction systems (A-B and A-C) were indirectly affected by the polyglutamine expansion region when going from the wild-type to the mutant. The un-common atom-atom pairs are not directly associated with the additional residues of the mutant (Fig. 41). We additionally see that the A-B system's mutant attractive interactions were diminished over the wild-type strain with a concomitant increase in inter-atomic distances.

Table 42: Un-common Results for Huntingtin (A-B)

	WT	MUT	%
UL TOT:	-7.01E+00	-2.53E+00	64%
UCOUL TOT:	-3.90E+00	-3.34E+00	14%
DISTANCE (AVG):	5.07E-01	6.03E-01	16%
UN-COMMON ATOM-ATOM PAIRS:	62	31	

Table 43: Un-common Results for Huntingtin (A-C)

	WT	MUT	%
UL TOT:	-8.13E+00	-2.91E+00	64%
UCOUL TOT:	-3.49E+00	-4.38E+00	20%
DISTANCE (AVG):	5.21E-01	6.06E-01	14%
UN-COMMON ATOM-ATOM PAIRS:	76	39	

The A-C pair was more varied in that its wild-type had better Lennard-Jones potentials and distance, but the mutant version had stronger Coulombic interactions.

Finally we turn our attention to the B-X family of interactions which for this case, only contained one member: B-C. The overall results for B-C are as follows,

Table 44: Fine Parsing Results for Huntingtin (B-C)

	WT	MUT	%
UL TOT:	-6.85E-01	-3.08E+00	78%
UCOUL TOT:	-9.37E+00	-1.03E+01	9%
DISTANCE (AVG):	7.86E-01	6.32E-01	20%

As can be seen, the mutant version of the protein has considerably better attractive interactions than the wild-type, specifically with respect to the Lennard-Jones potential and distance. Of interest was the dramatic increase (78%) of binding due to Lennard-Jones interactions for the mutant which was higher than any of the Lennard-Jones potential increases observed in the A-X family.

Looking at the Fine Parsing image results, we can see that the mutant strain has greater numbers of atom-atom interaction partners compared to the wild-type. In addition, these binding atoms are even more densely arranged for the mutant than for the wild-type which is (at least from a qualitative perspective) consistent with the data from Table 44.

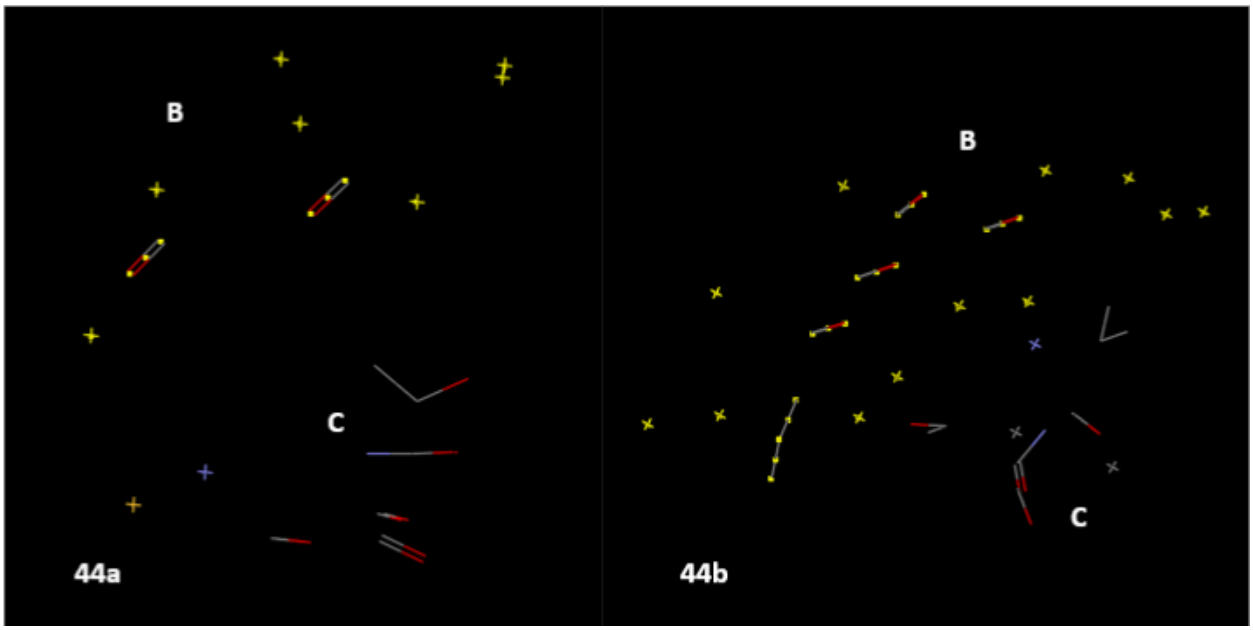


Fig. 44: Huntingtin wild-type (44a) and mutant (44b) Fine Parsing, B-C

Wild-type: 372-383 (B chain), 371-374 (C chain)

Mutant: 372-384 (B chain), 371-377 (C chain)

Coarse parsing image results show an overall retention of structure when going from the wild-type to the mutant. Detailed inspection of the two structures, however, shows that the mutant strain has slightly developed more structure than the wild-type. This development of structure is indicative of greater stability for the mutant which is hinted at in Table 44 by the increase in Lennard-Jones interactions and a decrease in average inter-atomic distance.

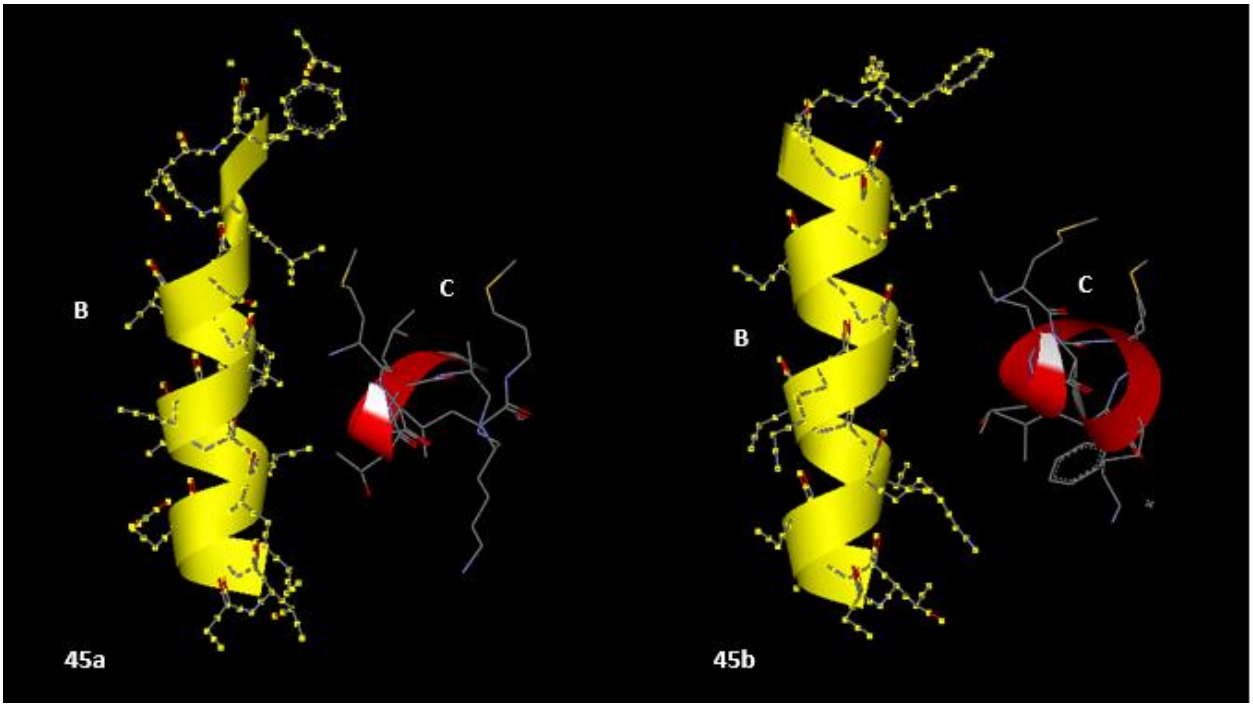


Fig. 45: Huntingtin wild-type (45a) and mutant (45b) coarse parsing, B-C

Wild-type: 372-391 (B chain), 371-381 (C chain)

Mutant: 372-391 (B chain), 371-381 (C chain)

The results for the common atom-atom pair interactions likewise indicated that the mutant strain had somewhat stronger attractive interactions, although the distances changed negligibly (Table 45).

Table 45: Common Results for Huntingtin (B-C)

	WT	MUT	%
UL TOT:	-3.47E-01	-4.88E-01	29%
UCOUL TOT:	-7.38E+00	-8.69E+00	15%
DISTANCE (AVG):	7.77E-01	7.36E-01	5%
COMMON ATOM-ATOM PAIRS:	21		

With the un-common interactions, it was observed that the Lennard-Jones potential increased when going from the wild-type to the mutant quite dramatically (by 87%) and the Coulombic potential decreased, but not as sharply (only 21%).

Table 46: Un-common Results for Huntingtin (B-C)

	WT	MUT	%
UL TOT:	-3.37E-01	-2.60E+00	87%
UCOUL TOT:	-2.00E+00	-1.58E+00	21%
DISTANCE (AVG):	8.04E-01	5.54E-01	31%
UN-COMMON ATOM-ATOM PAIRS:	10	28	

The dramatic increase in the Lennard-Jones potential for the mutant, however, was also accompanied by a decrease in average inter-atomic distance.

6.0 RESULTS AND DISCUSSION

This study proved very insightful in terms of the behavior that many seemingly un-related protein systems exhibited. As it turned out (and this occurred purely by coincidence), the six systems could be broken down into three main groups (where each group contained two protein systems), where each group was defined by a certain set of characteristics. Those three groups were originally identified by whether or not the biological activity of the protein was affected by the mutation(s). Other criteria were soon discovered to adhere to a trend marked by characteristics such as acquisition of additional secondary structure by the mutant and number of chains of the biologically active protein. It is important to note that although these observations

may hint at a potential classification system for intra-protein behavior, the study of additional systems is necessary to determine if classification can indeed be obtained. These observations are shown in Table 47.

Table 47: Potential PPI Organization Scheme

		Biology affected?	Mutant gains secondary structure?	Strain favored by potential shift?	Number of chains?	Total number of ligands?
CLASS1	TRANS	No	No	Mutant	4	3
	SOD1	No	Yes	Neither	2	2
CLASS2	DHFR	Yes	Yes	Mutant	1	2
	HRAS	Yes	Yes	Mutant	1	2
CLASS3	P53	Yes	No	Wild Type	4	1
	HUNT	Unknown	No	Wild Type	3	Unknown

These results were compiled from the overall Fine Parsing analyses of each system. For ease of comparison, these specific Fine Parsing results are in the appendix, Figs. B1-B6. Thus there may be a paradigm that unifies protein systems of diverse functions which may provide a rational means for studying these systems and developing therapeutics and treatments for them, but again, further studies on many more systems would be necessary before any such trends can be declared.

7.0 FUTURE WORK

In light of the observations summarized in Table 47, it would be prudent and interesting to test further cases that fall into one of the three general classes outlined. Do all one-chain protein systems whose mutants exhibit acquired secondary structure have the mutant strain

favored with higher Lennard-Jones and Coulombic values? Will such systems always have their biological function impaired by the mutation(s)? These are questions which can be undertaken as a direct "spin-off" from the results of the research reported in this work.

Additionally, one can extend further studies that make use of dynamic molecular simulations and models to further investigate why the observations reported in this work are so in the first place. With that research question in mind, it is important to remember that although OpenContact© is a powerful tool to begin dissecting and investigating proteins, the results it renders are purely from a static paradigm that do not take protein dynamics into consideration. As has been reported for proteins such as p53, many systems' biological activity rely on the ability to open up to accommodate ligands. Unfortunately, one can miss out on a lot of these conformational changes when dynamic considerations are neglected.

Lastly, now that theoretical work has been conducted on the select systems discussed in this paper, perhaps the next level of inquiry could involve devising methods to detect these systems in vivo. Perhaps a protocol that makes use of bio-mimetic peptides or aptamers can be developed to indicate if these aberrant proteins are present, and even to indicate the degree of aberrancy (a mutated p53 binds poorly to its ligand, this is not necessarily a guarantee of no binding whatsoever). In short, novel, additional technologies can potentially be developed that may advantage of any substantiated trends observed in later studies.

REFERENCES

1. Andersen D. Cystic Fibrosis of the Pancreas and Its Relation to Celiac Disease: A Clinical and Pathological Study. Am J Dis Child. 1938;56:344-399.

2. Ando Y, Jono H. Pathogenesis and Therapy for Transthyretin Related Amyloidosis. Rinsho Byori. 2008;56:114-120.

3. Auclair J, Brodkin H, D'Aquino AJ, Petsko G, Ringe D, Agar J. Structural Consequences of Cysteinylation of Cu/Zn-Superoxide Dismutase. Biochemistry. 2013;52:6145-6150.

4. Buseyne F, Janvier G, Teglas J, et al. Impact of Heterozygosity for the Chemokine Receptor CCR5 32-bp-Deleted Allele on Plasma Virus Load and CD4 T Lymphocytes in Perinatally Human Immunodeficiency Virus-Infected Children at 8 Years of Age. J Infect Dis. 1998;178:1019-1023.

5. Cario H, Smith D, Blom H, et al. Dihydrofolate Reductase Deficiency Due to a Homozygous DHFR Mutation Causes Megaloblastic Anemia and Cerebral Folate Deficiency Leading to Severe Neurologic Disease. The American Journal of Human Genetics. 2011;88:226-231.

6. Cell Signaling Technology. TTR (human). Available at: http://www.phosphosite.org/proteinAction.do;jsessionid=B65BFA5BEC8DF7B323D05F72F02FC5AE?id=3290854&showAllSites=true. Accessed 7/13, 2014.

7. Cell Signaling Technology. SOD1 (human). Available at: http://www.phosphosite.org/proteinAction.do?id=15569&showAllSites=true. Accessed 07/13, 2014.

8. Cell Signaling Technology. p53 (human). Available at: http://www.phosphosite.org/proteinAction.do?id=465&showAllSites=true. Accessed 07/13, 2014.

9. Cell Signaling Technology. HRas (human). Available at: http://www.phosphosite.org/proteinAction.do?id=2698&showAllSites=true. Accessed 07/13, 2014.

10. Cell Signaling Technology. DHFR (human). Available at: http://www.phosphosite.org/proteinAction.do?id=2888800&showAllSites=true. Accessed 07/13, 2014.

11. Cell Signaling Technology. Huntingtin (human). Available at: http://www.phosphosite.org/proteinAction.do?id=1292&showAllSites=true. Accessed 07/13, 2014.

12. Cendron L, Trovato A, Seno F, et al. Amyloidogenic Potential of Transthyretin Variants INSIGHTS FROM STRUCTURAL AND COMPUTATIONAL ANALYSES. JMB. 2009;284:25832-25841.

13. Clarke J, Mahuran D, Sathe S, et al. An Open-Label Phase I/II Clinical Trial of Pyrimethamine for the Treatment of Patients Affected with Chronic GM2 Gangliosidosis (Tay-Sachs or Sandhoff Variants). Molecular Genetics and Metabolism. 2004;102:6-12.

14. Cornell W, Cieplak P, Bayly C, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc. 1995;117:5179-5197.

15. Franken S, Scheidig A, Krengel U, et al. Three-Dimensional Structuresand Properties of a Transforming and a Nontransforming Glycine-12 Mutant of P21-HRas. Biochemistry. 1993;32:8411-8420.

16. Galaleldeen A, Strange R, Whitson L, et al. Structural and Biophysical Properties of Metal-Free Pathogenic SOD1 Mutants A4V and G93A. Arch Biochem Biophys. 2009;492:40-47.

17. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci Signal. 2013;6:pl1.

18. Goodsell D. p53 Tumor Suppressor. Available at: http://www.rcsb.org/pdb/101/motm.do?momID=31. Accessed 8/13, 2014.

19. Jorgensen W. The Many Roles of Computation in Drug Discovery. Science. 2004;303:1813-1818.

20. Kim M, Chelliah Y, Kim S, Otwinowski Z, Bezprozvanny I. Secondary Structure of Huntingtin Amino-Terminal Region. Structure. 2009;17:1205-1212.

21. Kim M. Beta Conformation of Polyglutamine Track Revealed by a Crystal Structure of Huntingtin N-Terminal Region with Insertion of Three Histidine Residues. Prion. 2013;7:221-228.

22. Kollman P, Massova I., Reyes C., et al. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. Acc Chem Res. 2000;33:889-897.

23. Krall A, Brunn J, Kankanala S, Peters M. A Simple Contact Mapping Algorithm for Identifying Potential Peptide Mimetics in Protein-Protein Interaction Partners. Proteins. 2014;Epub ahead of print.

24. Kuhn B, Kollman P, et al. Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities By a Combination of Molecular Mechanics and Continuum Solvent Models. Journal of Medicinal Chemistry. 2000;43:3786-3791.

25. Lefebvre P. Assembly and Motility of Eukaryotic Cilia and Flagella. Lessons from Chlamydomonas reinhardtii. Plant Physiol. 2001;127:1500-1507.

26. Malecka K, Ho W, Marmorstein R. Crystal Structure of a p53 Core Tetramer Bound to DNA. Oncogene. 2009;28:325-333.

27. Miyakawa T, Morikawa R, Takasu M, et al. Molecular Dynamics Simulations of the Hras-GTP Complex and the Hras-GDP Complex. Quantum Chemistry. 2013;113:2333-2337.

28. Monaco H. The transthyretin-retinol-binding protein complex. BBA. 2000;1482:65-72.

29. Mora P, Carbajo R, Pineda-Lucena A, Sanchez del Pino M, Perez-Paya E. Solvent-exposed residues located in the β-sheet modulate the stability of the tetramerization domain of p53—A structural and combinatorial approach. Proteins: Structure, Function, and Bioinformatics. 2008;71:1670-1685.

30. Muraoka S, Shima F, Araki M, et al. Crystal Structures of the State 1 Conformations of the GTP-Bound H-Ras Protein and its Oncogenic G12V and Q61L Mutants. FEBS Letters. 2012;586:1715-1718.

31. Niu Q, Ybe J. Crystal structure at 2.8Å of Huntingtin-interacting protein 1 (HIP1) coiled-coil domain reveals a charged surface suitable for HIP-protein interactor (HIPPI). Journal of Molecular Biology. 2008;375:1197-1205.

32. Oefner C, D'Arcy A, Winkler F. Crystal Structure of Human Dihydrofolate Reductase Complexed with Folate. Eur J Biochem. 1988;174:377-385.

33. Palinathan S, Mohamedmohaideen N, Snee W, Kelly J, Sacchettini J. Structural Insight into pH-Induced Conformational Changes within the Native Human Transthyretin Tetramer. JMB. 2008;382:1157-1167.

34. Peters M. Force descriptions. In: McCombs KP, ed. Real-Time Biomolecular Simulations. New York, New York: McGraw Hill; 2007:58-62.

35. Ramirez D, Mejiba Gomez S, Mason R. Mechanism of Hydrogen Peroxide-Induced Cu,Zn-Superoxide Dismutase-Centered Radical Formation as Explored by Immuno-spin trapping: the Role of Copper- and Carbonate Radical Anion-Mediated Oxidations. Free Radical Biology & Medicine. 2005;38:201-214.

36. Reynaud E. Protein Misfolding and Degenerative Diseases. Nature Education. 2010;3:28.

37. Selverstone Valentine J, Doucette P, Zittin Potter S. Copper-Zinc Superoxide Dismutase and Amyotrophic Lateral Sclerosis . Annu Rev Biochem. 2005;74:563-593.

38. Soltis DE, Soltis PS, Zanis MJ. Phylogeny of Seed Plants Based on Evidence From Eight Genes. American Journal of Botany. 2002;89:1670-1681.

39. Stobener K, Klein P, Reiser S, Horsch M, Kufer K, Hasse H. Multicriteria Optimization of Molecular Force Fields by Pareto Approach. Fluid Phase Equilibria. 2014;373f:100-108.

40. Ulrich E, Akutsu H, Doreleijers J, et al. BMRB Featured System: Dihydrofolate Reductase. Available at: www.bmrb.wisc.edu/featuredSys/dhfr/dhfr1.html. Accessed July/3, 2014.

41. Volpato J, Yachnin B, Blanchet J, et al. Multiple Confomers in Active Site of Human Dihydrofolate Reductase F31R/Q35E Double Mutant Suggest Structural Basis for Methotrexate Resistance. J Biol Chem. 2009;284:20079-20089.

APPENDIX A

Table A1 (pdb structure file anatomy):

1	2	3	4	5	6	7	8	9	10	11	12
ATOM	7	N	PRO	A	11	36.535	13.092	36.537	1.00	32.680	N
ATOM	8	CA	PRO	A	11	35.653	12.829	35.409	1.00	31.150	C
ATOM	9	C	PRO	A	11	36.326	12.719	34.040	1.00	29.600	C
ATOM	10	O	PRO	A	11	35.659	12.367	33.077	1.00	28.290	O

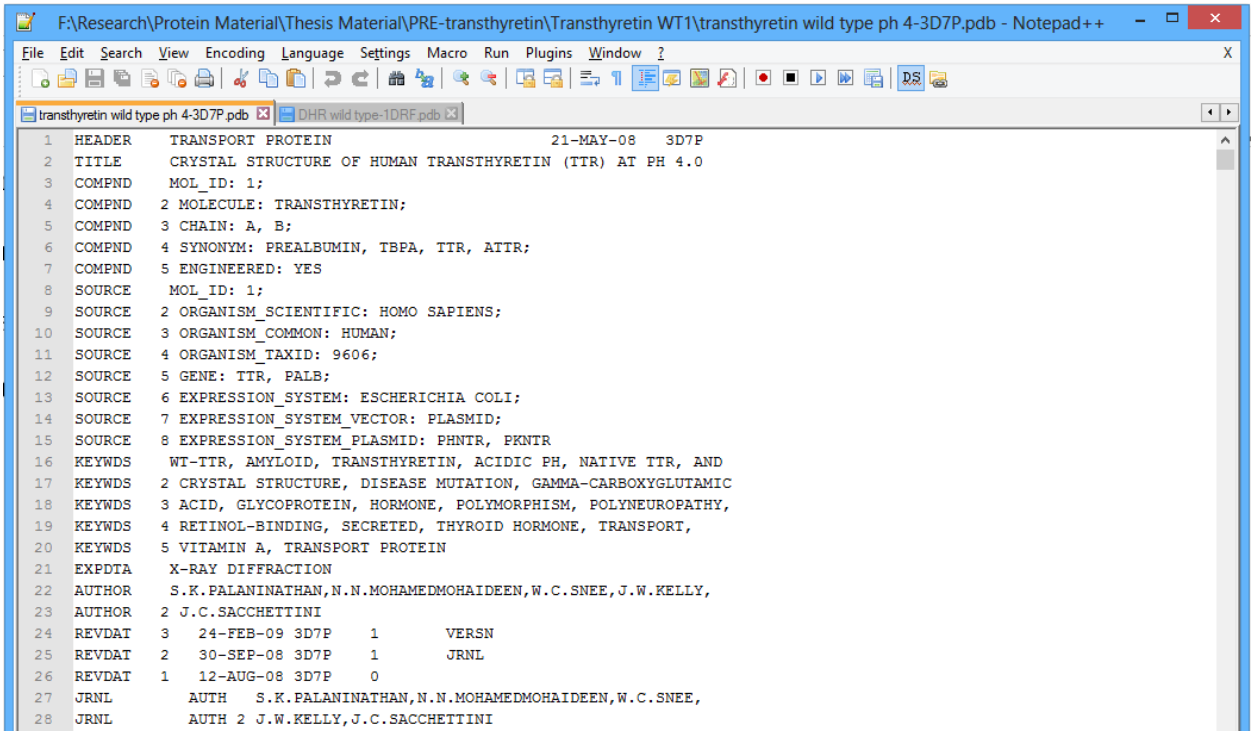


Fig. A1: Information section of a typical pdb file

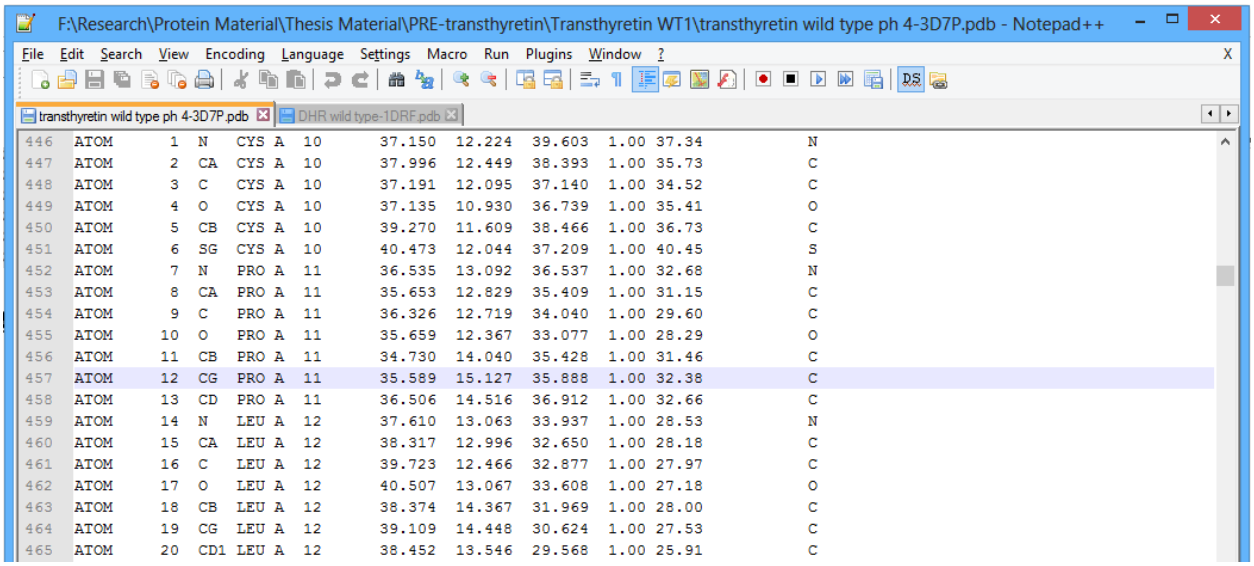


Fig. A2: Coordinate and atom data section (what DS Viewer Pro plots)

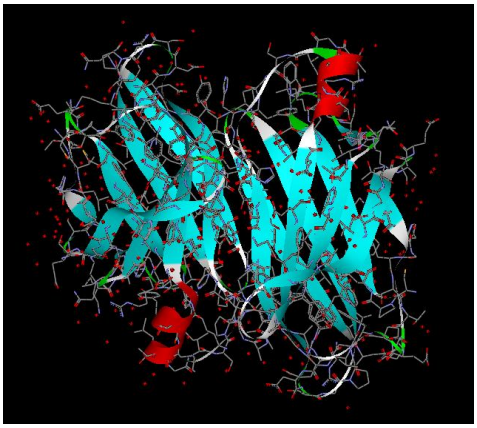


Fig. A3: Flat ribbon transthyretin with overlaid atoms shown in wire-frame schematic

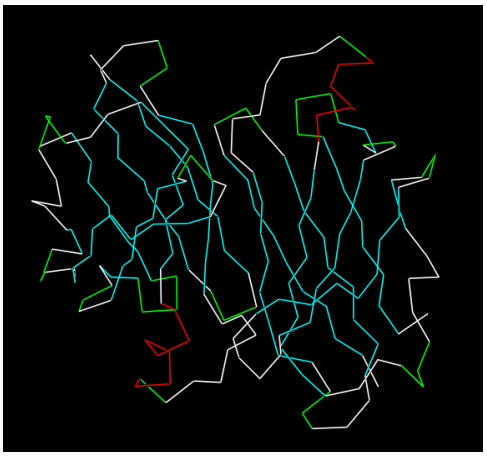


Fig. A4: Transthyretin in wire schema

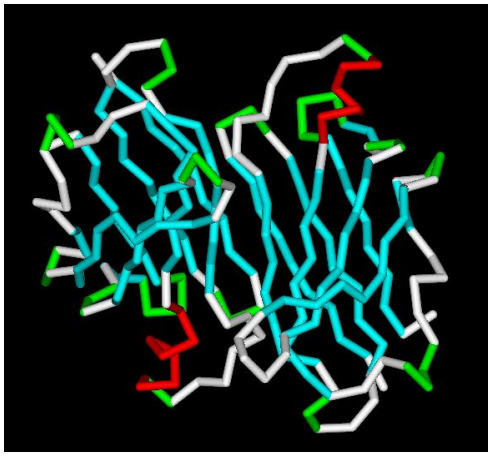


Fig. A5: Transthyretin in stick schema

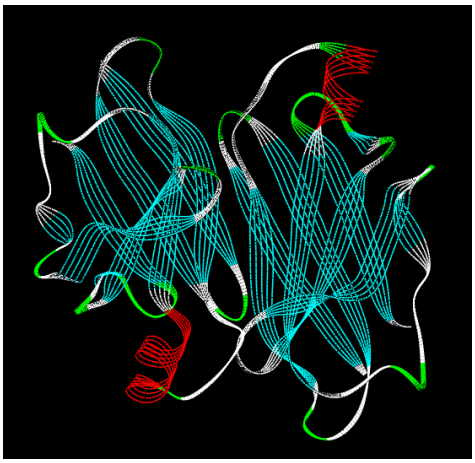


Fig. A6: Transthyretin in line ribbon schema

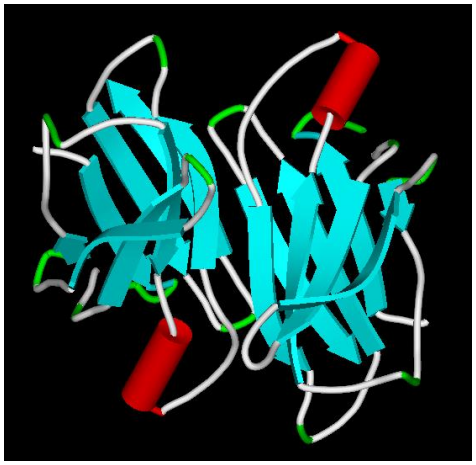


Fig. A7: Transthyretin in schematic schema

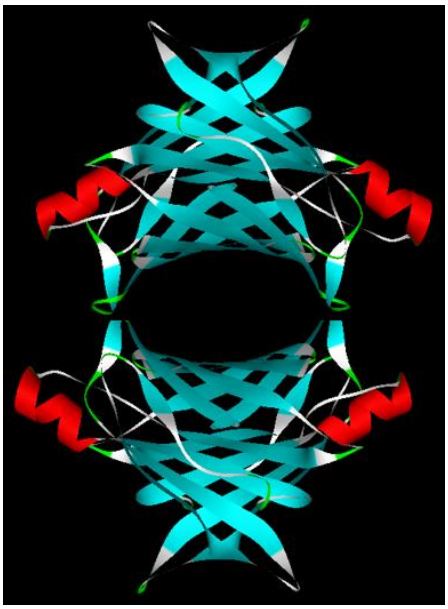


Fig. A8: Tetramer of transthyretin (biologically active tetramer)

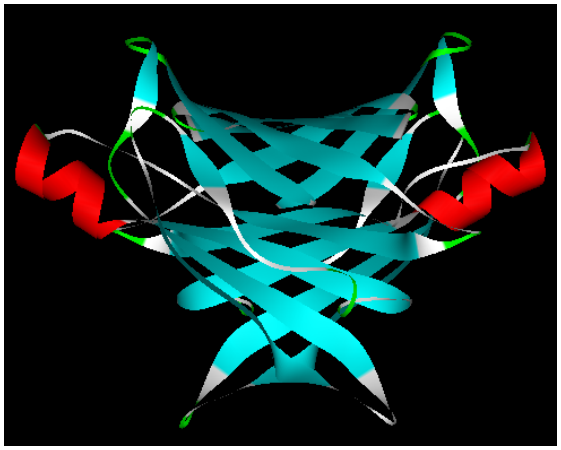


Fig. A9: Dimer of transthyretin

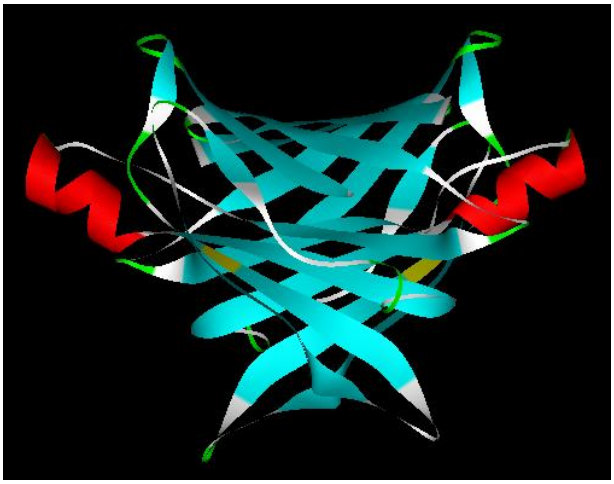


Fig. A10: Mutated residues hi-lighted in transthyretin

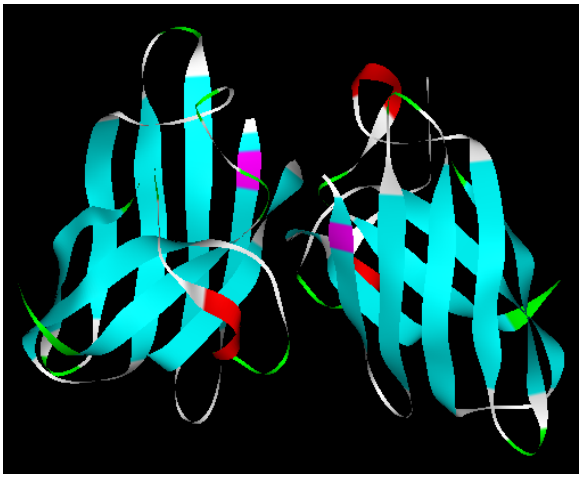


Fig. A11: Mutated residues hi-lighted in

SOD1

APPENDIX B

B1:

p53 Symmettry:

A-B

	WT	MUT	%
UL TOT:	-5.35E+01	-4.25E+01	21%
UCOUL TOT:	-3.04E+02	-2.60E+02	14%
DISTANCE (AVG):	5.89E-01	6.08E-01	3%

C-D

	WT	MUT	%
UL TOT:	-5.92E+01	-5.31E+01	10%
UCOUL TOT:	-2.79E+02	-2.81E+02	1%
DISTANCE (AVG):	5.82E-01	5.91E-01	2%

A-C

	WT	MUT	%
UL TOT:	-1.56E+00	-1.10E+00	30%
UCOUL TOT:	-2.66E+01	-3.59E+01	26%
DISTANCE (AVG):	8.04E-01	8.03E-01	0.10%

B-D

	WT	MUT	%
UL TOT:	-2.48E+00	-1.25E+00	50%
UCOUL TOT:	-2.70E+01	-1.18E+01	56%
DISTANCE (AVG):	7.68E-01	7.81E-01	2%

A-D

	WT	MUT	%
UL TOT:	-5.31E+00	-5.40E+00	2%
UCOUL TOT:	-6.87E+01	-7.81E+01	12%
DISTANCE (AVG):	6.97E-01	7.11E-01	2%

B-C

	WT	MUT	%
UL TOT:	-3.74E+00	-7.15E+00	48%
UCOUL TOT:	-5.46E+01	-8.33E+01	34%
DISTANCE (AVG):	7.24E-01	6.91E-01	5%

B2:

Huntingtin Symmetry:

A-B

	WT	MUT	%
UL TOT:	-7.83E+00	-3.31E+00	58%
UCOUL TOT:	-1.32E+01	-1.24E+01	6%
DISTANCE (AVG):	5.57E-01	6.42E-01	13%

A-C

	WT	MUT	%
UL TOT:	-8.20E+00	-3.05E+00	63%
UCOUL TOT:	-8.86E+00	-1.11E+01	20%
DISTANCE (AVG):	5.68E-01	6.45E-01	12%

B-C

	WT	MUT	%
UL TOT:	-6.85E-01	-3.08E+00	78%
UCOUL TOT:	-9.37E+00	-1.03E+01	9%
DISTANCE (AVG):	7.86E-01	6.32E-01	20%

B3:

Transthyretin:

	WT	MUT	%
UL TOT:	-5.41E+01	-6.36E+01	15%
UCOUL TOT:	-2.38E+02	-2.46E+02	4%
DISTANCE (AVG):	6.07E-01	5.82E-01	4%

B4:

SOD1:

	WT	MUT	%
UL TOT:	-3.99E+01	-4.07E+01	2%
UCOUL TOT:	-2.34E+02	-2.11E+02	10%
DISTANCE (AVG):	6.55E-01	6.43E-01	2%

B5:

DHFR:

	WT	MUT	%
UL TOT:	-1.31E+02	-1.53E+02	14%
UCOUL TOT:	-4.56E+02	-4.77E+02	4%
DISTANCE (AVG):	5.87E-01	5.79E-01	1%

B6:

HRAS:

	WT	MUT	%
UL TOT:	-8.02E+01	-9.07E+01	12%
UCOUL TOT:	-3.81E+02	-4.98E+02	23%
DISTANCE (AVG):	6.09E-01	6.06E-01	0.50%